Spatial Epidemiology of Tuberculosis in Hong Kong

PANG, Tak Ting Phoebe

A Thesis Submitted in Partial Fulfilment

of the Requirements for the Degree of

Master of Philosophy

in

Public Health

The Chinese University of Hong Kong

September 2010

Thesis /Assessment Committee

Professor Benny Chung Ying ZEE (Chair)
Professor Shui Shan LEE (Thesis Supervisor)
Professor Tung FUNG (Thesis Co-supervisor)
Professor David Shu Cheong HUI (Committee Member)
Dr Poh Chin LAI (External Examiner)

Acknowledgement

I would like to express sincere gratitude to my supervisor Professor LEE Shui Shan (Professor of Stanley Ho Centre for Emerging Infectious Diseases, the Chinese University of Hong Kong) for offering me an opportunity to work on this subject and for his guidance through the three-year study. I would also like to thank my co-supervisor Professor FUNG Tung (Professor of Department of Geography and Resource Management, the Chinese University of Hong Kong) for his guidance and invaluable comments.

Special thanks to Dr LEUNG Chi Chiu (Consultant Chest Physician, Department of Health) for his support in TB data collection, continuous guidance and invaluable discussions. Moreover, I would also like to thank the staff in TB and Chest Clinic for their assistance.

I am grateful to Mr Joe Ng from CentaMap for his assistance so that geocoding becomes possible in this study. Particular thanks are extended to Miss Denise Chan, Mr Yhon Lam, and Miss Brenda Lee from the Stanley Ho Centre for Emerging Infectious Diseases, for their help and encouragement.

Abstract of thesis entitled

Spatial epidemiology of tuberculosis in Hong Kong

Submitted by PANG, Tak Ting, Phoebe

for the degree Master of Philosophy in Public Health

at The Chinese University of Hong Kong in June 2010

Tuberculosis (TB) is a public health problem worldwide. Its resurgence could be attributed to the emergence of HIV/AIDS epidemic and drug resistance, and also explained by various social determinants. In light of the relatively minimal impact of HIV/AIDS and drug resistance on TB in Hong Kong, this study aimed at evaluating the influence of social determinants on local epidemiology.

A spatial approach using centrographic analysis, exploratory spatial data analysis (ESDA), ordinary linear regression (OLR) and geographically weighted regression (GWR) was adopted to describe the spatial distribution and identify significant neighbourhood determinants of TB. Over 17 000 TB cases notified from 2005 to 2007 were geocoded down to the District Council Constituency Area (DCCA) level. DCCA-based TB data were standardized by underlying population structure and the standardized TB ratio was correlated with neighbourhood determinants which were categorized in five domains.

The standardized TB ratio demonstrated a spatially varied pattern as illustrated by ESDA. TB ratio was significantly associated with all determinants in univariate analysis but only marital status, place of birth, population density and low household income remained independent predictors in OLR model. These determinants were further put into a GWR model. Significant spatial variation of the association between TB and

determinants were uncovered and model improvement of GWR over OLR was noted.

Despite the virtual elimination of absolute poverty in Hong Kong, low household income, a surrogate of poverty, demonstrated the strongest association with TB ratio among other significant determinants in the OLR model. However, GWR model was able to identify specific DCCAs that TB ratio was predominantly modulated by non Hong Kong born and population density instead.

The study helps to enhance our understanding of the TB epidemiology in Hong Kong from a spatial perspective and this would enable a better allocation of resources for TB control.

摘要

肺結核是一個全球公共衛生議題。全球疫情發展除了受到愛滋病病毒及耐藥性結核的影響外，同時也受到不少社會因素影響。有見於愛滋病病毒及耐藥性結核對香港的肺結核疫情相對比較輕微，故本研究目的旨在探討目前各種危險因素對香港的肺結核疫情的影響。

本研究嘗試以地理的分析架構，利用空間中心統計、探索性空間數據分析方法 (ESDA)、傳統線性迴歸 (OLR) 再配合地理加權迴歸 (GWR)，來解析肺結核在香港的地理分佈及識別主要影響肺結核的危險因素。此研究收集了由二零零五年至零七年超過一萬七千多個肺結核呈報病例作爲肺結核數據。利用地理編碼方法，肺結核數據以區議會選區聚合作爲計算單位，並利用各選區的人口分佈計算標準化肺結核呈報率 (TB SNR)。利用迴歸方法分析 TB SNR 與各項同樣以區議會選區作爲計算單位的危險因素之關係。

ESDA 結果顯示 TB SNR 表現出顯著的空間變化。單變量分析數據指出 TB SNR 與所有危險因素有關，而 OLR 則歸納出四個與 TB SNR 有關的危險因素，包括：婚姻狀況、出生地點、人口密度及低收入家庭。隨後，GWR 除了提升了回歸模型的闡釋力，其結果更顯示出各因素與 TB SNR 關係存在空間變化。

本研究不但呈現肺結核空間分佈的特徵，同時也反映香港小區危險因素與肺結核的空間關係，有助相關機構更有效地投放疾控資源。

List of Figures

List of Tables

Abbreviation

AIC	Akaike Information Criterion
AIDS	Acquired immunodeficiency syndrome
BCG	Bacille Calmette-Guérin
DC	District Councils
DCCA	District Council Constituency Areas
DOTS	Directly observed treatment (short course)
ESDA	Exploratory spatial data analysis
GIS	Geographic information system
GPS	Global positioning system
GWR	Geographically weighted regression
HIV	Human immunodeficiency virus
LISA	Local indicator of spatial association
MAUP	Modifiable area unit problem
MDR-TB	Multidrug-resistant tuberculosis
OLR	Ordinary linear regression
SDE	Standard deviational ellipse
SMO	Survey and Mapping Office
SNR	Standardized notification ratio
TB	Tuberculosis
TB SNR	Tuberculosis standardized notification rate
TPU	Tertiary Planning Units
WHO	World Health Organization
XDR-TB	Extensively drug-resistant tuberculosis

Table of Contents

CHAPTER ONE INTRODUCTION

1.1 Historical perspective of tuberculosis

Tuberculosis (TB) is an ancient disease. Pathological signs of tubercular decay discovered in Egyptian mummies proved the existence of human TB since prehistoric era (Donoghue et al, 2004). In the 18th century, the first TB epidemic occurred in Europe as a result of industrialization. Cites were being rapidly developed and urban settlements were improperly arranged. People were thrown together in squatters where housing and hygiene condition was poor. Crowding living environments, poor hygiene and people suffering from malnutrition constituted ideal circumstances for the accelerated spread of TB. Following the global diffusion of industrialization and population migration, TB epidemic swept through the world in the next 150 years (Stead, 1997). TB

caused tremendous number of morbidity and mortality before effective treatment to the disease was available.

1.1.1 Sanatorium care

Later in mid 1800s, a botanist introduced sanatorium care and it was soon widely adopted as the first specific treatment for TB. Infected people were sent to sanatorium where they were provided with plenty of food, spacious living area and fresh air. The improved living environment allowed patients to strengthen host immunity to fight against the TB bacteria in their body, and gradually led to recovery. Despite the progress and improvement was slow, TB was subsequently brought into control (Wilson, 1990). Until today, the rationale of improving one's living condition and ensuring nutrition intake remains a core control strategy in defending disease progression from infection.

1.1.2 Vaccination

Medical care in sanatorium had been used for almost 100 years as a primary treatment to TB until the discovery of *mycobacterium tuberculosis,* the bacillus causing the disease, by a German microbiologist Robert Koch in 1882 (Sakula, 1982). Koch has excited the world by being able to isolate and identify the bacteria using a special staining technique. With the ability to see the bacteria, rapid development in vaccine and drugs took place in the 19th century. A French bacteriologist Calmette, together with veterinarian called

Guerin, developed the Bacille Calmette-Guérin (BCG) vaccine using attenuated bacterium (Andersen & Doherty, 2005). The vaccine was found efficacious in protecting young children from having tuberculous meningitis. This first BCG vaccine was used in human in 1912 and is still widely used today.

1.1.3 Drug treatment

The beginning of chemotherapy era to TB treatment was marked by the successful isolation of streptomycin, the first effective antibiotic against the bacteria, in 1943 (Mitchison, 2005). The emergence of resistant strain resulted from the monotherapy using one drug urged more effective anti-TB drugs to be subsequently developed in the next few years. A combination usage of drugs in treatment regimen was later found effective to suppress drug resistance. Chemotherapy using four drugs including isoniazid, rifampicin, pyrazinamide and either ethambutol or streptomycin becomes a global standard regimen for TB treatment.

1.1.4 Transmission dynamics of tuberculosis

The devastating effect of TB to the vast amount of population is manifested by its nature of disease transmission. TB is an airborne infectious disease. The ongoing chain of transmission in population could be characterized as the most distinctive feature of infectious disease epidemiology. The bacteria causing TB can be transmitted in the air from a person with active pulmonary

TB to others by coughing or sneezing. If not treated, an infectious person could infect 10 to 15 people on average every year (World Health Organization, 2009). Exposure to the bacteria may lead to infection. However, most individuals infected with TB would develop an asymptomatic latent infection when the host immunity helps suppressing the multiplication of TB bacteria in the body. Over the life course in healthy individuals, approximately 10% of the latent infections would progress to active TB disease, either in form of progressive primary infection (referring to disease occurring within 5 years of infection), endogenous reactivation of the exiting infection (referring to disease with onset 5 years or more after infection), exogenous re-infection with a new TB strain (referring to first disease episode within 5 years of reinfection) (Vynnycky & Fine, 1997). The risk for progression to TB disease is markedly higher for people with immunosuppressive diseases, such as diabetes mellitus and human immunodeficiency virus (HIV) / acquired immunodeficiency syndrome (AIDS). For people co-infected with HIV/AIDS and TB, the risk for developing TB disease is high, at a rate of 5% to 10% per year (Selwyn, 1992; Moreno et al, 1993; Jansa et al, 1998).

1.1.5 Resurgence of tuberculosis

As contact between human being is simply a necessary part of life, TB was spread in a high speed and was the most fatal infectious disease worldwide. Later, vaccination and chemotherapy, in addition to the improved living

condition, accelerated the decline of TB morbidity and mortality and the epidemic was finally under control. Chemotherapy is effective not only in reducing TB morbidity and mortality, but also the infectiousness of TB patient soon after they began treatment. However, inappropriate prescription of drugs, as well as failure in ensuring treatment adherence, leads to the development of drug resistant strain of TB. The emergency of drug resistant TB has alarmed the world as it is difficult to treat. The global TB epidemic was further worsened by the emergence and spread of HIV/AIDS epidemic. As a result, TB is, again, threatening the world.

1.2 Current global and local tuberculosis epidemiology

The resurgence of TB in various parts of the world can be partially attributed to two factors that are commonly identified as the underlying reasons for the global emergence and re-emergence: HIV/AIDS and TB drug resistance.

1.2.1 Tuberculosis and HIV/AIDS, drug resistance in the world

HIV/AIDS has a profound effect on TB. The immune system of people living with HIV/AIDS is seriously impaired. The weakened immunity of the host is no longer able to defend against the dormant TB bacteria, as a result latent TB infection would be more likely to progress to active disease. In 2008, 15% (n =1.4 million) of the global new TB cases were HIV positive (World Health Organization, 2009). Seventy-eight percent of global TB-HIV co-infection cases in the world were found in Africa (Figure 1.1). The intersection of HIV/AIDS and TB is expanding the existing pool of TB cases, as it would precipitate disease progression and recurrence.

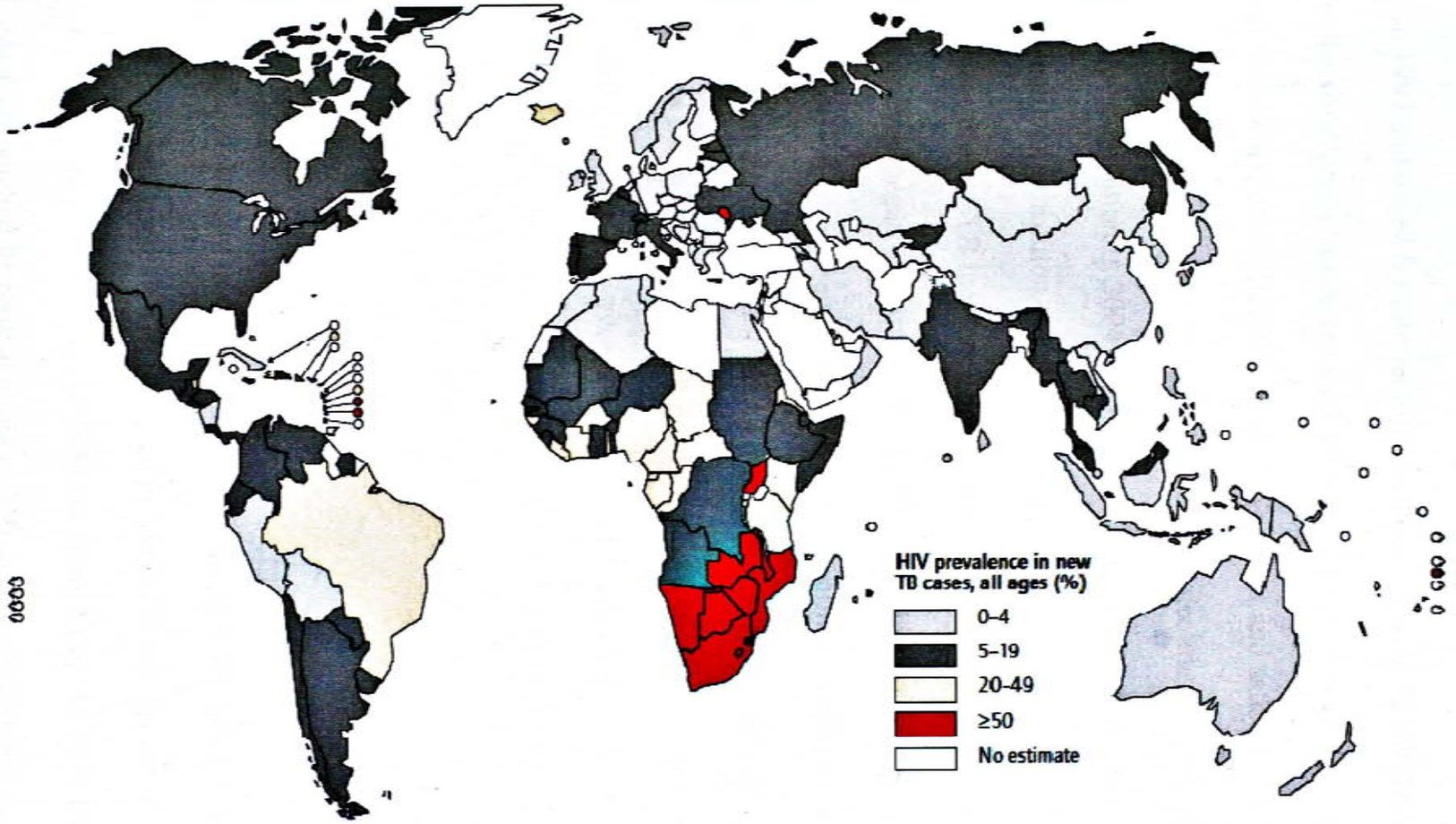
Due to the long course of TB treatment, non-adherence and emergence of drug resistance were reported ever since the earliest days of chemotherapy. The emergence of drug resistance TB resulting from inadequate course of treatment turns TB into a more dangerous disease. Drug resistance TB can be classified into two types, namely multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB). MDR-TB is caused by bacteria that are resistant to the most effective anti-TB drugs (isoniazid and rifampicin).

MDR-TB may develop either from primary infection or in the course of the treatment. XDR-TB is a form of TB caused by bacteria that are resistant to isoniazid and rifampicin (i.e. MDR-TB) as well as any fluoroquinolone and any of the second-line anti-TB injectable drugs (amikacin, kanamycin or capreomycin). Since 1994, 114 countries have reported cases of MDR-TB and 58 countries have reported at least one case of XDR-TB as of March 2010 (World Health Organization, 2010).

If drug resistant TB is left untreated, there would be ongoing transmission of infections that are drug resistant, which are very difficult and expensive to treat. Moreover, the situation has been worsened by the development of XDR-TB, which is virtually untreatable using current medicines.

Figure 1.1 Map showing the percentage of HIV-positive in new TB cases in 2008

A total of 1.4 million TB/HIV co-infection cases were reported in 2008 and 78% of them were found in Africa.



Source: World Health Organization, 2009

1.2.2 Global epidemiology of tuberculosis

In light of the re-emergence of TB over the world, WHO declared TB as a global emergency in 1993 and urged for international collaborations in fighting against this curable disease. Various measures were advocated to combat this global epidemic but the strengthened effort has not led to prompt and satisfactory ramification. Though the global incidence rate started to decline from 142 new cases per 100,000 population in 2004 to 137 in 2007, the decline has been very slow (World Health Organization, 2009). Still, TB is a worldwide pandemic. In some countries the drop in TB incidence began to level off in the mid 1980s and then stagnated or even began to rise. In 2009, more than 2 billion people, one third of the world population were infected with TB and 1.8 million people were killed by this disease. From the map of estimated TB incidence (Figure 1.2), the highest TB incidence rates are found in African countries, where HIV/AIDS epidemic is causing an explosion of TB. However, we should not overlook the Asia region (the South-East Asia and Western Pacific regions), where it accounts for 55% new TB cases in the world.

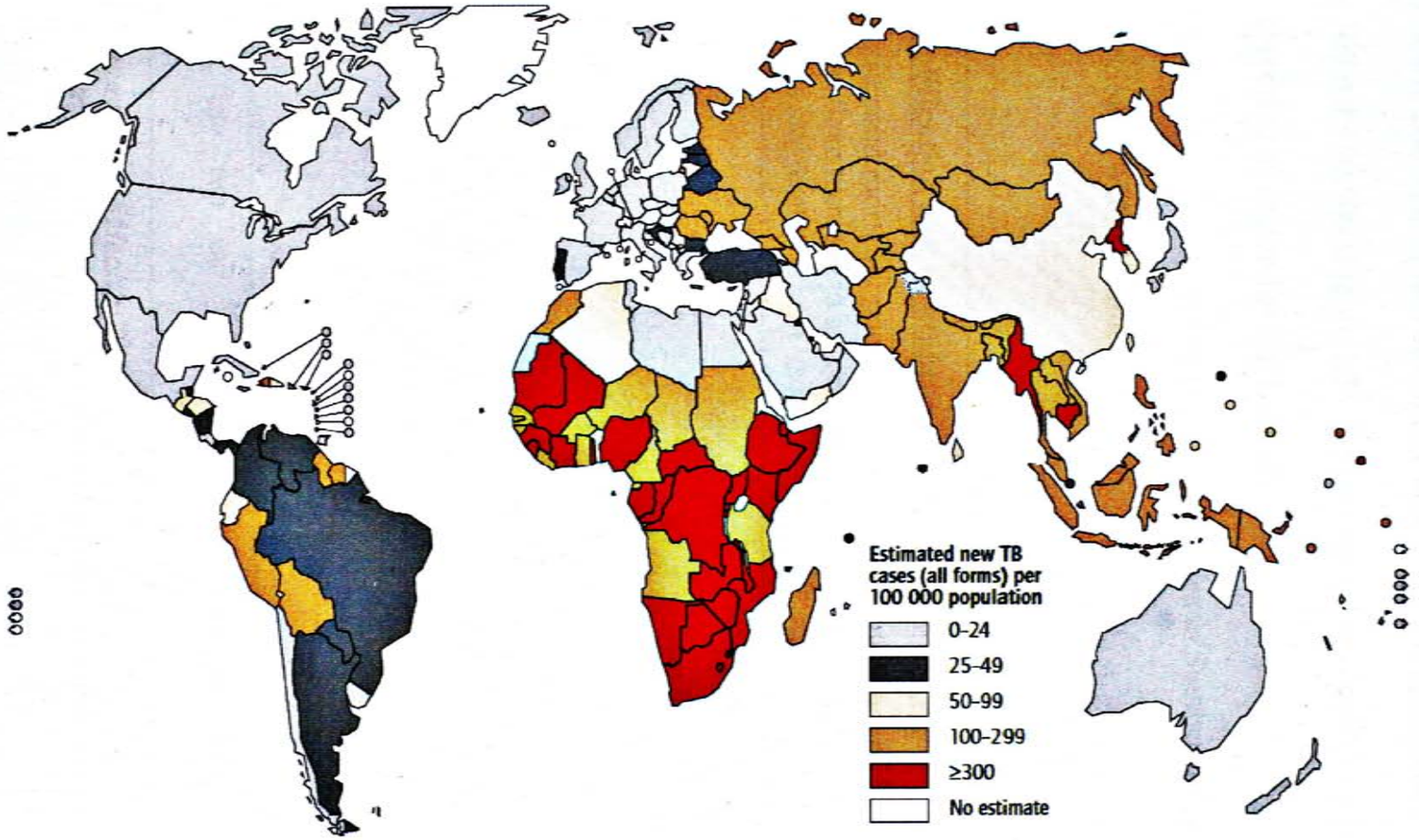
1.2.3 Local epidemiology of tuberculosis

As a metropolitan city located in the Southeast Asia, Hong Kong had the first case of phthisis, the historical name of TB, recorded in 1849. TB was the most fatal disease during the war periods in 1940s to 1950s in Hong Kong. In 1939 the mortality rate of TB reached its peak at a rate of 250 deaths per 100,000 population and TB became a notifiable disease in the same year. The TB notification system is

maintained by Tuberculosis and Chest Service unit, Department of Health, Hong Kong. The notification system is core to the local TB control programme as it is a surveillance system which ensures statutory ongoing collection of data that are essential to understand local epidemiology and implement control measures. In 1970, a standardized 6-month directly observed treatment short-course (DOTS) chemotherapy to all patients was implemented, subsequently both TB morbidity and mortality were drastically reduced. The treatment success rate is about 80% at one year and about 84.8% at two years counting from date of starting anti-TB treatment (Tam et al, 2003).

Figure 1.2 Map showing the estimated incidence rate of new TB cases in 2008

The estimated global incidence rate was 139 cases per 100000 population in 2008 and the estimated incidence rate in African countries was a double to that in Southeast Asia Region, which accounts for one-third of the world TB cases.

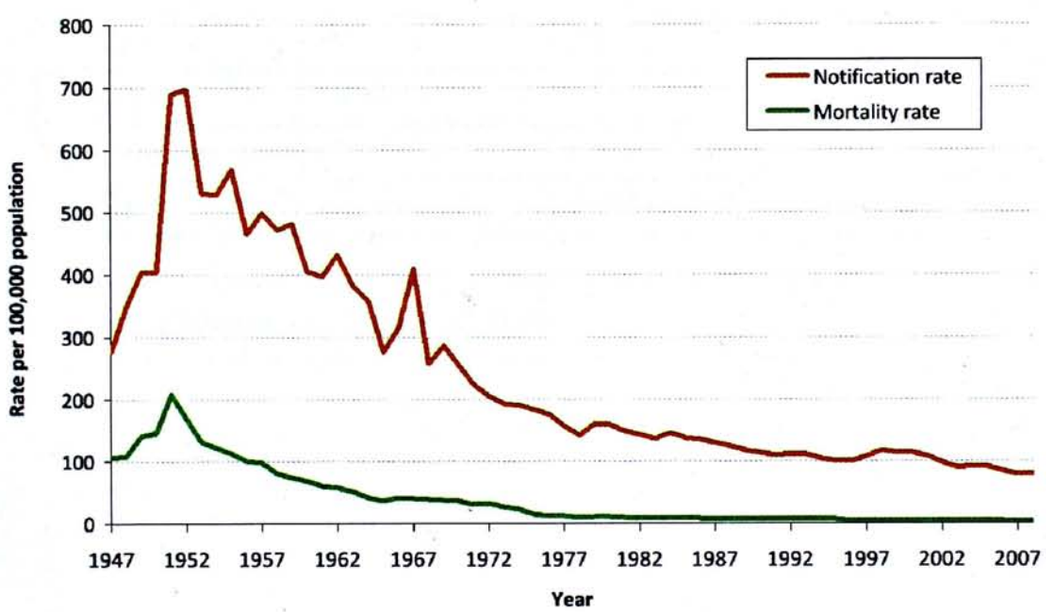


Source: World Health Organization, 2009

Notification rate of TB in Hong Kong has shown an overall downward trend in the past 50 years. The incidence rate decreased from a peak of 697 per 100,000 in 1952 to 78.9 per 100,000 in 2007. In 2002 it was the first time for TB notification rate dropped below 100 per 100,000 (Figure 1.3). Despite the trend is declining Hong Kong has been classified as region having intermediate burden of TB with good health infrastructure by WHO, comparing to other countries in Southeast Asia and Western Pacific regions. (World Health Organization, 2005)

Figure 1.3 TB notification and mortality rate 1947 – 2008

Both TB notification and mortality rate have been declining in the past 50 years. While mortality has been reduced to a very low level, the declining trend of notification rate was stagnant in recent years.



Data Source: Department of Health, 2008

The declining rate is much slower than expected and fluctuations did occur from time to time. In 2008, TB notification rose slightly to 80.8 per 100,000 and a total of 5635 TB cases were notified (Department of Health, 2008). Among the notified cases, 64% were male and 41% were aged 65 or above (Table 1.1). Since Hong Kong population is aging, elderly TB cases are likely to reflect the TB burden associated with the waning immunity with age.

Table 1.1 Table showing the age and sex distribution of TB case in Hong Kong, 2008

A total of 5635 TB cases were notified in Hong Kong. A majority of cases were male and 41% were elderly.

Age group	Male	Female	Total
0 to 14	17	31	48 (1%)
15 to 64	1904	1367	3271 (58%)
65 or above	1707	609	2316 (41%)
Total	3628 (64%)	2007 (36%)	5635 (100%)

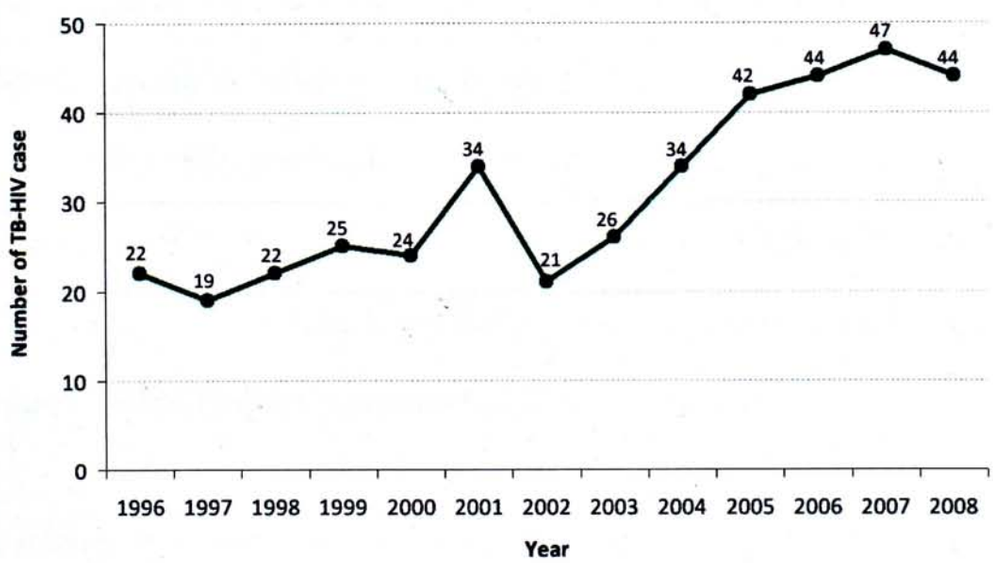
Data Source: Department of Health, 2008

1.2.4 Tuberculosis, HIV/AIDS and drug resistance in Hong Kong

Currently in Hong Kong, the prevalence of HIV infection remains low, with a case count of about 3000. According to an unlinked anonymous screening for HIV in TB & Chest Service in 2008, a total of 44 cases of TB/HIV co-infection has been reported (Department of Health, 2009). Although the number of TB-HIV cases reported demonstrates a steady increase over years, the total number of co-infection cases remains at low level (Figure 1.4). TB/HIV co-infection contributes a very low proportion of the total TB caseload of which less than 1% of TB patients were HIV seropositive.

Figure 1.4 Number of TB-HIV cases reported to the TB-HIV Registry 1996 – 2008

Despite the increasing number of TB/HIV co-infection cases, less than 1% of total TB cases were co-infection.



Data Source: Department of Health, 2009

With the effective implementation of DOTS and DOTS-plus in Hong Kong, the overall drug resistance problem is under progressive control. Currently, MDR-TB accounts for 1% of the total TB cases (Department of Health, 2008). However the global emergence of MDR- and XDR-TB are posing increasing difficulties in the control of TB locally, especially in view of the frequent outbound population movement and high rates of drug-resistant TB in some of our neighbouring areas including China and Thailand. In general, the impact of HIV/AIDS and drug resistance on local epidemiology has been, though minimal, of rising concern.

1.2.5 Approaches in studying tuberculosis epidemiology

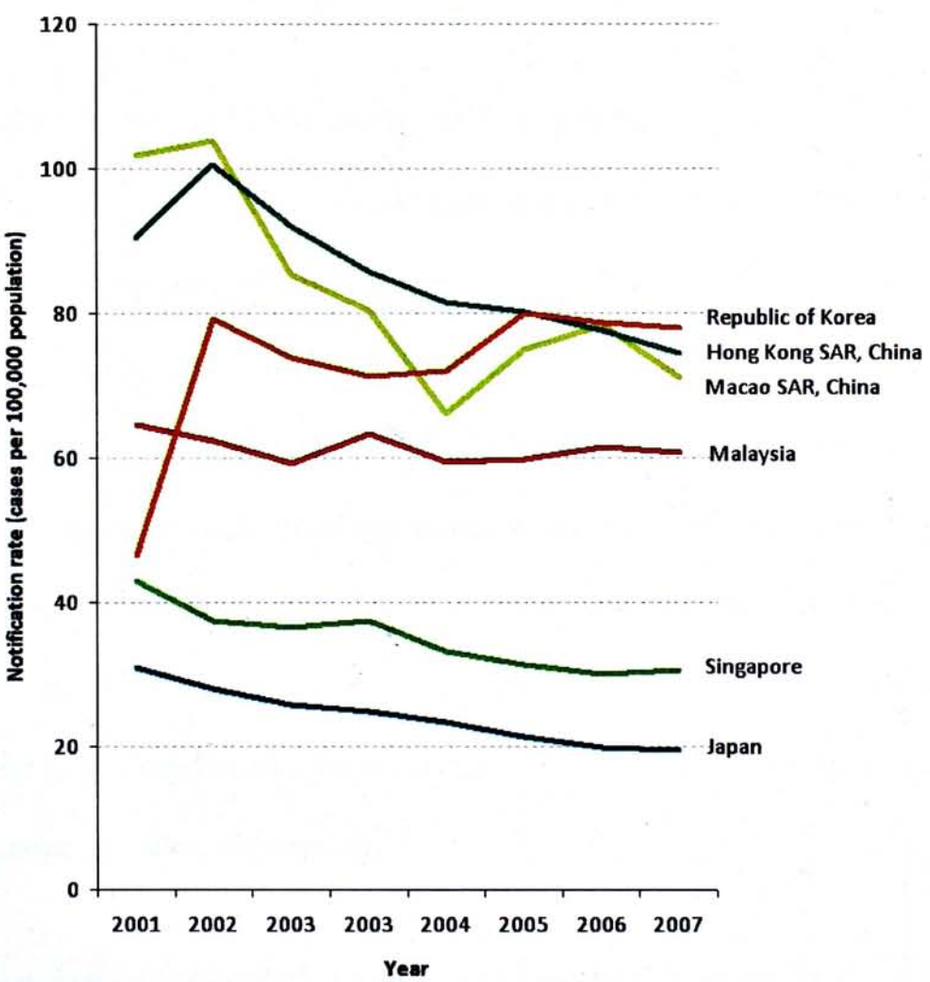
Obviously TB is a complex problem and the observed pattern cannot be fully explained by HIV/AIDS and drug resistance. The insignificant decline in incidence rate continues to put TB a public health burden not only to Hong Kong. As shown in Figure 1.5, an insignificant decreasing trend has also been observed in some neighbouring countries which are also classified as countries having intermediate TB burden. While HIV and MDR-TB have contributed to the resurgence of TB in most part of the world, the disease is far from under control in countries like Singapore, Japan, Vietnam and Hong Kong where both factors are relatively uncommon (World Health Organization, 2005).

While some researchers focus on biomedical investigation such as vaccine development, others try to understand the re-emergence of TB from a public health perspective. Despite biomedical interventions remains the core strategy in the

current control paradigm, evidence from historical practice in controlling TB by sanatorium care, improved housing condition and provision of nutrition, demonstrated the importance of underlying determinants of TB, of which can be social (Lönnroth et al, 2009), psychological (Dye et al, 2009) and environmental (Lienhardt, 2001).

Figure 1.5 Comparison of trends of notification rates of TB, intermediate burden countries

Regions with medium level of TB were classified as intermediate burden countries by Western Pacific Regional Office of WHO. As shown in the figure, the declining trend becomes stable or even reverses in recent years.



Data Source: World Health Organization, 2009

1.3 Determinants of tuberculosis epidemiology

Studying determinants of a disease, or commonly called risk factors, is a typical epidemiology investigation. By definition, epidemiology is concerned about disease and also all health-related outcome among populations, rather than individuals. This unique approach in studying human health makes epidemiology different from conventional medicine and constitutes the basic science of public health. In studying TB, epidemiologists describe observed TB pattern in terms of frequency and distribution, and relate the observed pattern to factors that may lead to TB infection and/or disease.

1.3.1 TB determinants in the triad of epidemiology

By using various techniques, epidemiologists distinguish TB as well as its determinants in terms of *time*, *person* and *place*, the longstanding triad of epidemiology. Investigation of TB outbreak (Lofy et al, 2006; Ohkado et al, 2009) and seasonal pattern (Leung et al, 2005; Luquero et al, 2008) are examples of focus on the *time* component. Epidemiology studies focusing on the difference between sex (Hudelson, 1996), age groups (Wu et al, 2008) and race (Serpa et al, 2009) are examples to illustrate the *person* component. Studies on detection of disease cluster (Nunes, 2007), and modelling of disease spread (de Vries et al, 2008) are examples that emphasize the *place* component.

Interest in the effects of *place* and environment on TB is not new. However, with the technological and scientific constraints, a majority of studies highlighting TB

and *place* are primarily descriptive. On the other hand, *person* becomes the core focus of most TB epidemiology studies, with the upsurge of biostatistics. TB patients are compared with healthy people, in terms of the demographic and social characteristics. Most of these studies do not take into the account of the spatial variability of the determinants in relation to TB, and have assumed that the identified determinants pose equal effect on population subgroups though they are living in different areas.

The many facets of TB epidemiology are complicated. Determinants of TB can be universal in some cases, for example, the emergence of resistance; it can also be *place-specific*, for example, the transmission dynamics in different settings. In fact, different risk factors indentified in geographically-diverse literature demonstrated the presence of heterogeneity of risk factors in different regions. This spatial heterogeneity proves the place-specific nature of risk factors and, this heterogeneity does not exist only in county level but also in neighbourhood, or even smaller, level.

1.3.2 Rise of spatial epidemiology

Although the *place* dimension is an inherent part of the conceptual framework of epidemiology, it is only recently that they have been incorporated explicitly into ecological theory, sampling design and modelling techniques. The reason is mostly due to the unavailability of technical packages/tools for researchers to statistically quantify the spatial difference and association.

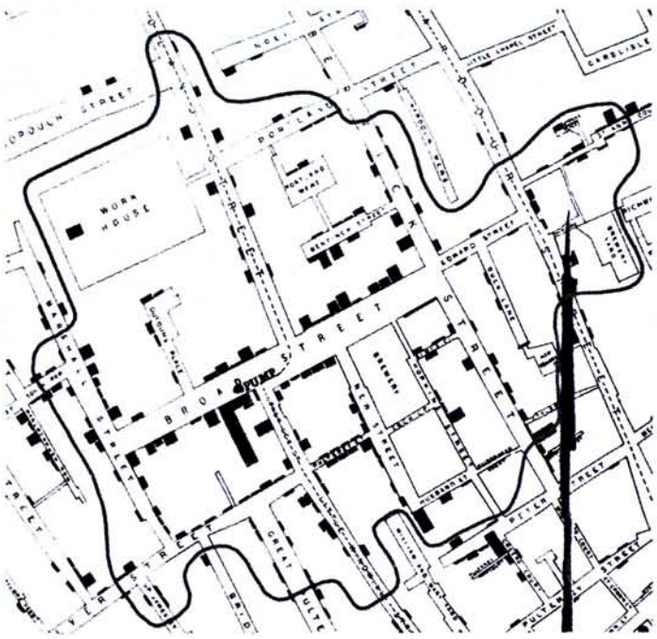
The very first novel application of spatial analysis of an infectious disease is often attributed to John Snow when he was studying the cholera outbreak in London in 1854. To substantiate his hypothesis of the water-borne nature of cholera, he gathered residential data of the deaths and marked each death on an existing map of area affected by the outbreak. To represent where multiple deaths occurred he used a single bar, so that the density of deaths at individual locations would be evident. The map of the apparent concentration of deaths around a single water pump justified John Snow's argument against an airborne origin for the spread of cholera generally. Furthermore, he manually drew a thiessen polygon (Figure 1.6), a kind of proximity analysis, to illustrate how majority of deaths were close to the infected Broad Street pump than to any other water pumps. As a result, he successfully demonstrated the associations between cholera deaths and water pump in a striking spatial distribution.

Centuries passed, technological advance gives rise to a growing number of tools are now available to address spatial questions. On top of the availability of tools, growing awareness of the aggregated effects of neighbourhood determinants on health is now coming up with individual effects. There is increasing acceptance of evidence that people's area of residence may influence their health either in addition to or in interaction with their individual characteristics. Sometimes intra-individual variability makes a single measure a poor marker for the person's exposure (e.g. a cluster of unemployment may have a different effect of individual unemployment). Most of the time environmental determinants could not be conceptualized in individual level. In this case, it is necessary to aggregate disease

pattern and suspected determinants by different places, treating spatial units, rather than individuals, as the basic analytical unit of the study.

Figure 1.6 John Snow and his cholera map

By drawing the thiessen polygon, John Snow was able to illustrate how the majority of death cases were closer to the infected Broad Street pump.



Source: The John Snow Archive and Research Companion http://johnsnow.matrix.msu.edu/images/online_companion/chapter_images/fig12-6.jpg. Accessed on 26 May 2010.

The development of spatial statistical methods, availability of spatial data, advances in geographic information system (GIS) and the significance in examining the aggregated effect of environment on health provide a new opportunity to study epidemiology by combining epidemiology, statistics and GIS. This emerging field of analysis is called *spatial epidemiology* (Elliot & Wartenberg, 2004).

1.4 Recent developments of spatial epidemiology

Spatial epidemiology refers to the description and analysis of geographic variation in disease with respect to demographic, environmental, behavioural, socioeconomic, genetic risk factors. The core idea of spatial epidemiology is based on the *Tobler's First Law of Geography* (Tobler, 1970). In 1970, Tobler stated that "Everything is related to everything else, but near things are more related than distant things". Therefore, in the analytical framework of spatial epidemiology, disease is not evenly distributed across *space*. The resultant disease pattern could be influenced by the varying distribution of risk factors over *space*.

1.4.1 Spatial epidemiology and infectious disease

Spatial epidemiology is of particular significance in studying infectious disease. As infectious disease must be transmitted via direct contact, the presence of disease in certain area may imply possible episodes of transmission. There is growing number of studies applying spatial analytics in studying infectious disease. In a special report, Brownstein illustrated how the HealthMap system, as well as cartographic visualization of the disease, contributes for situational awareness in the global evolution of 2009 H1N1 influenza pandemic (Brownstein, 2010). In a study conducted in Russia on HIV/AIDS, Heimer was able to identify a missing link between injection drug users and HIV prevalence by using spatial analysis (Heimer, 2008). A local study conducted from Lai on SARS in Hong Kong has demonstrated clearly how spatial analysis is useful in the detection of disease clusters (Lai, 2004).

1.4.2 Disease mapping

From the classic study of cholera to recent analyses on emerging infectious diseases, cartographic visualization of disease pattern has always been the first and fundamental session of methodology. Disease maps have been used since 1800 when thematic cartography began. The first recorded disease map was a dot map showing cases of yellow fever in New York around 1800 (Walter, 2001). Despite the usefulness of disease maps, they were usually applied as an incidental supplement rather than an essential part of an epidemiology report or a research study. However, in the past two decades, disease mapping are more frequently adopted, as a result of development in computer cartography and GIS.

1.4.3 Geographic information system

While maps remain core to spatial analysis, a good deal of the spatial analytical tools has been paralleled by the development of GIS. By definition, a GIS is a computer-based system for collecting, editing, integrating, visualizing and analysing spatial data. GIS first emerged in the 1960s, but it has evolved rapidly in the past 25 years and only relatively recently that geographers have begun to explore these developments. In addition to its mapping functionalities, GIS offers a powerful platform for database management. In a GIS environment, data from various sources could be linked, maintained and analysed under a relational database structure. In a geodatabase, all data must be spatially referenced, either to the global coordinate system or to a local projected coordinate system. Georeferencing of residential address collected from patient or study population

could be done via the use of standardized geocoding machine or Global Positioning System (GPS), a tool of rising popularity.

1.4.4 Statistics in spatial epidemiology

Increasing availability of spatially-referenced health data and population data encourages the development of statistical techniques to spatial epidemiology. Spatial statistics is different from conventional statistics in terms of their basic assumption. While conventional statistics assume data independence, spatial statistics emphasize data dependence and measure how dependent the data is.

Disease and ill health always have a spatial dimension. Over a century ago, epidemiologists and public health practitioners were already using maps to help assimilating and understanding the spatial dynamics of disease. Adopting *Tobler's First Law of Geography* as the principle, sophisticated techniques of spatial analysis and GIS continues to be developed to extract explanatory and predictive power from space. At the same time it is expected to see a growing number of studies using such a cross disciplinary approach to revisit the old concept of epidemiology.

CHAPTER TWO LITERATURE REVIEW

2.1 Objective of literature review

In order to distinguish possible research gap and formulate the hypothesis in this study, a literature review was conducted

1. to explore how spatial perspective of TB epidemiology has been addressed
2. to identify neighbourhood determinants associated with TB incidence
3. to examine the techniques used in spatial studies of TB

2.2 Literature search

In this section, the process of literature search as well as the general characteristics of studies included is discussed.

2.2.1 Strategy for literature search

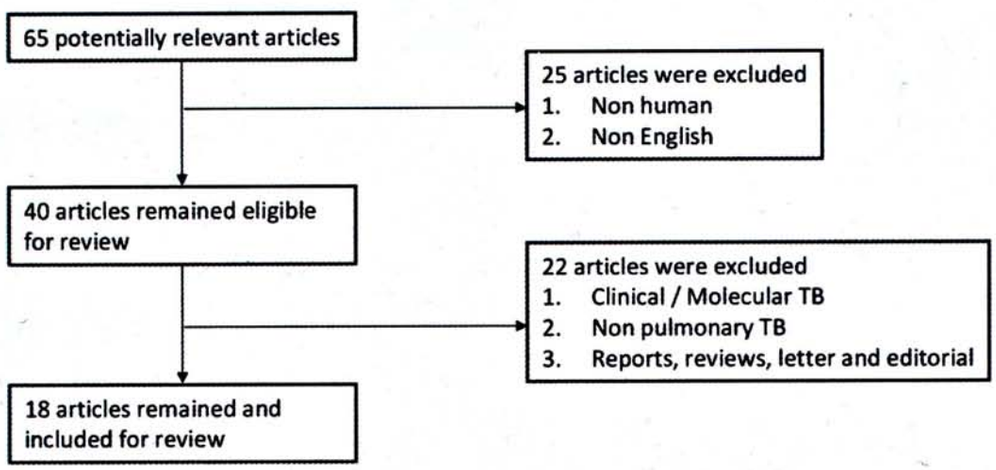
A literature search was conducted in February 2010 using PubMed, an electronic database. To ensure studies searched did examine explicitly the spatial perspective of TB epidemiology, title of articles must contain search terms "tuberculosis" and either "spatial", "geographic" or "geographical". There is no restriction on publication date. The search was limited to publications in English and studies conducted on human. Articles on clinical and molecular aspect of TB, non pulmonary TB and articles that were national reports, review papers, letters and editorials were excluded. Only original articles are included.

2.2.2 Results for literature search

The search in PubMed returned a total of 65 hits. After excluding 25 articles that studied TB in animals and used language other than English, 40 articles remained. Among the 40 articles, 22 of them were further excluded as they were about clinical and molecular TB (n = 10), extra-pulmonary TB (n = 10), and national reports, review article, letter and editorials (n = 2). As a result, 18 articles remained eligible and were subjected to complete review. Process of literature search is illustrated in Figure 2.1.

Figure 2.1 Flow of literature search

The literature search was conducted in February 2010, using an online database PubMed. Search words included tuberculosis, spatial, geographic and geographical. As a result, 65 articles were found and 18 of them were eligible for a complete review.



A literature code starting with a letter "L" was assigned to each article. For convenience, the literature code is used for in-text referencing in this chapter. The corresponding literature code for a particular article is shown in Table 2.1.

Table 2.1 List of studies reviewed and the corresponding code

Studies were assigned a code starting with a letter "L"

Code	Author	Title of article
L01	Randremanana et al.	Spatial clustering of pulmonary tuberculosis and impact of the care factors in Antananarivo City 2009
L02	Uthman et al.	Spatial and temporal variations in incidence of tuberculosis in Africa, 1991 to 2005 2008
L03	Jia et al.	Spatial analysis of tuberculosis cases in migrants and permanent residents, Beijing, 2000-2006 2008
L04	Higgs et al.	Early detection of tuberculosis outbreaks among the San Francisco homeless: trade-offs between spatial resolution and temporal scale 2007
L05	Haase et al.	Use of geographic and genotyping tools to characterize tuberculosis transmission in Montreal 2007
L06	Onozuka et al.	Geographic prediction of tuberculosis clusters in Fukuoka, Japan, using the space-time scan statistic 2007
L07	Tiwari et al.	Investigation of geo-spatial hotspots for the occurrence of tuberculosis in Almora district, India, using GIS and spatial scan statistic 2006
L08	Chan-yeung et al.	Socio-demographic and geographic indicators and distribution of tuberculosis in Hong Kong: a spatial analysis 2005
L09	Yeh et al.	Incidence of tuberculosis in mountain areas and surrounding townships: dose-response relationship by geographic analysis 2005
L10	Jacobson et al.	Changes in the geographical distribution of tuberculosis patients in Veracruz, Mexico, after reinforcement of a tuberculosis control programme 2005
L11	Munch et al.	Tuberculosis transmission patterns in a high-incidence area: a spatial analysis 2003
L12	Kistemann et al.	Spatial patterns of tuberculosis incidence in Cologne (Germany) 2002
L13	Bishai et al	Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy 1998
L14	Davidow et al.	Geographic diversity in tuberculosis trends and directly observed therapy, New York City, 1991 to 1994 1997
L15	Beyers et al.	The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high-incidence community 1996
L16	Weber et al.	Epidemiology of tuberculosis in North Carolina, 1966 to 1986: analysis of demographic features, geographic variation, AIDS, migrant workers, and site of infection 1989
L17	Froggatt et al.	Tuberculosis: spatial and demographic incidence in Bradford, 1980-2 1985
L18	Yanagawa et al.	Geographical pattern of tuberculosis and related factors in Japan 1981

Table 2.2 shows the characteristics of reviewed articles. The 18 studies were conducted in 13 countries. Six articles were conducted in Asia, 6 were conducted in Americas, 4 were conducted in Africa and 2 were conducted in Europe. According to the World Bank classification system, 10, 5 and 2 studies were conducted in high income economies, middle income economies and low income economies respectively. As Taiwan is not a member state of United Nations, no classification was applied. Two thirds (n = 12) of the study were published after 2000 and one-sixth (n = 3) of them were published before 1990s. All studies obtained TB data from the national surveillance system and all TB subjects included in these studies were TB disease case. A maximum number of 84366 TB cases were reported from the study conducted in Taiwan over a study period of 7 years. Large sample size of TB cases is essential to ensure statistical stability when cases were aggregated into spatial units.

The geographic distribution of the articles reveals research concern over 4 continents, implying the global attention to TB epidemiology. A larger proportion of studies were conducted in developed and high income countries, which could be due to the availability of spatial data. Successful mapping and application of spatial analysis heavily depends on availability of accurate spatial data. The relatively more established surveillance systems in developed countries ensure stable and reliable supply of disease data with locational information. On the other hand, obtaining regularly updated spatial information of disease and population data in developing countries is one of the main obstacles in introducing spatial analyses in

health as if requires enormous effort and investment in building national spatial databases.

Table 2.2 Studies included in the literature review

A total of 18 studies were identified. These studies were conducted in four world regions including Asia, Americas, Europe and Africa.

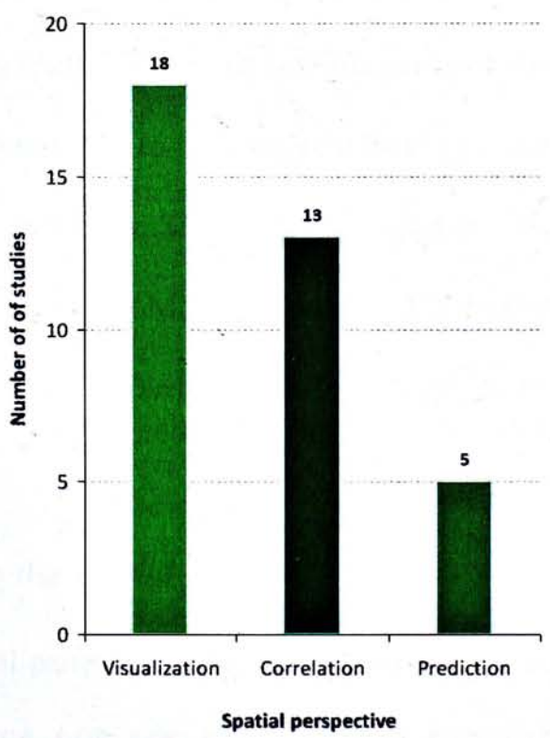
Code	Author	Publication year	Country	Country classification #	Study period (year)	TB cases
L01	Randremanana et al.	2009	Madagascar	Low income economies	3	2161
L02	Uthman et al.	2008	53 African countries	Low income economies	15	*
L03	Jia et al.	2008	China	Low-middle income economies	7	23026
L04	Higgs et al.	2007	USA	High income economies	12	392
L05	Haase et al.	2007	Canada	High income economies	5	863
L06	Onozuka et al.	2007	Japan	High income economies	6	9119
L07	Tiwari et al.	2006	India	Low-middle income economies	3	1635
L08	Chan-yeung et al.	2005	Hong Kong	High income economies	3	2332
L09	Yeh et al.	2005	Taiwan	N/A	7	84366
L10	Jacobson et al.	2005	Mexico	Upper middle income economies	5	436
L11	Munch et al.	2003	South Africa	Upper middle income economies	6	717
L12	Kistemann et al.	2002	Germany	High income economies	12	2903
L13	Bishai et al	1998	USA	High income economies	2.5	182
L14	Davidow et al.	1997	USA	High income economies	4	*
L15	Beyers et al.	1996	South Africa	Upper middle income economies	10	4011
L16	Weber et al.	1989	USA	High income economies	21	21115
L17	Froggatt et al.	1985	UK	High income economies	3	444
L18	Yanagawa et al.	1981	Japan	High income economies	1	80629
# Country classification based the measurement of GNI per capita by World Bank						
* No exact number of TB case was indicated in the study						

2.3 Spatial perspective in tuberculosis epidemiology

The common goal in all studies was to investigate the observed spatial difference in TB distribution. The investigation was performed from three different perspectives including 1) visualization of spatial distribution, 2) correlation of spatial pattern to hypothesized determinants, and 3) prediction of spatial distribution. Figure 2.2 shows the number of studies addressing TB epidemiology from different perspectives. The number of studies involving correlation and prediction is expected to decrease as they require for more sophisticated GIS programmes and spatial analytical techniques.

Figure 2.2 Three perspectives in investigating a spatial pattern

All studies visualized TB pattern as the fundamental aim in the study. While 13 of them continued to pursuit correlation analysis of the observed pattern, 5 of them tried to model and predict the TB pattern with its determinants.



2.3.1 Mapping the spatial pattern

Being fundamental and essential in all spatial analyses, visualization of TB pattern was regarded as the primary research objective in all reviewed studies (Figure 2.2). A pattern is a distinctive form that can be detected and described, of which disease mapping a way to detect and describe the spatial pattern of TB by researchers through linking observations and knowledge. Mapping TB distribution could help identifying the overall spatial pattern. Twelve studies examined spatial pattern through mapping various measures of TB, including case number (L03, L04, L07, L08, L13, L17), rates (L09, L12, L14, L15, L16, L18) and relative risk (L01). These maps were analysed visually and supplemented with a textual description which highlights interesting pattern of TB distribution.

With the advent of statistical packages, researchers became able to describe spatial pattern in a standard way. It has therefore given rise to a growing number of studies aiming at illustrating a spatial pattern as specific areas of elevated risk. Hotspots and clusters are common terms to refer to areas having a unexpected number of cases located in close proximity. Six studies included mapping of clusters of which clusters, which were defined by Moran's Index (L02), Scan Statistics (L06), K-function (L11) and distance buffering (L05, L09, L10).

2.3.2 Understanding the spatial pattern

Identification of spatial pattern is an important step to advance research hypotheses. Thirteen (L01, L03, L05, L08, L09, L10, L11, L12, L13, L15, L16, L17, L18) out

of 18 studies continued their description of spatial pattern with correlation analysis (Figure 2.2), which aimed at explaining the observed spatial distribution with possible risk factors. As a result, a list of biologic, socioeconomic and environmental and care factors were found to be associated with TB. The spatial distribution of these factors was likely to explain much of the spatial variation of TB epidemiology in the study area. Though causative relationships could hardly be determined by correlation analysis, the approach could uncover significant association and help formulating hypothesis for further exploration. More detailed discussion on determinants of TB is presented in the next section.

2.3.3 Modelling the spatial pattern

Five (L01, L08, L11, L12, L18) out of the 13 studies exploring correlation attempted to model the spatial distribution of TB with significant factors and to predict local spread of TB (Figure 2.2).The risk factors were further shortlisted according to their significance in contributing to the prediction of TB. With the prediction models, the diversity and magnitude of risk factors to TB were revealed. Prediction of TB pattern with significant determinants was found particularly useful for national surveillance of TB epidemiology (L18), so that health officials could be updated with determinants of TB which might be subject to changes from time to time (L08, L12) and to encourage area-specific prevention and control programme (L01, L11).

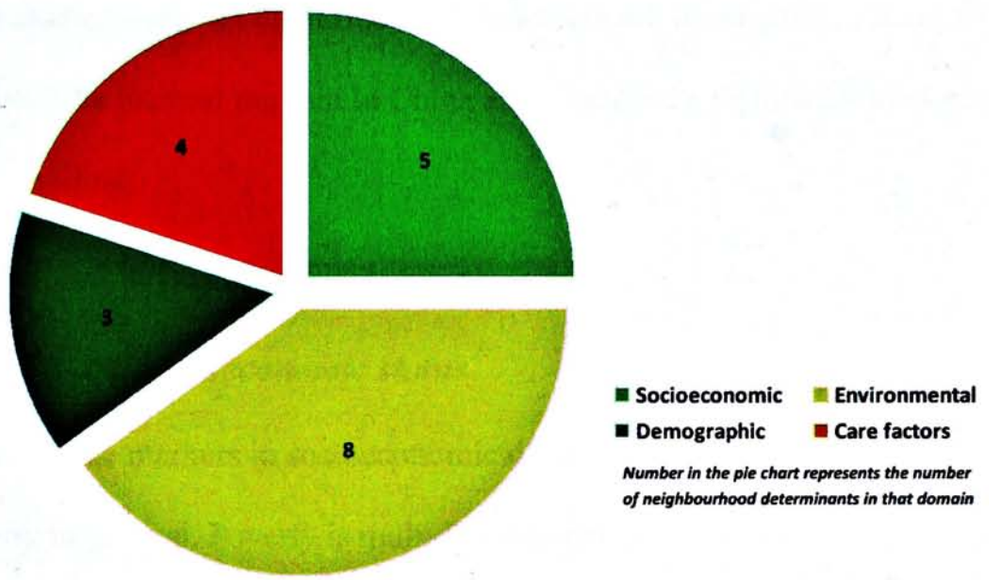
2.4 Neighbourhood determinants of tuberculosis

All articles reviewed are ecological studies and spatial units were considered the basic unit of analysis. Determinants (or risk factors) were aggregated into spatial units. Since the aggregated determinants were used to represent the environmental characteristics of where people reside and was assumed to be most influential to one's health, these determinants were usually referred as neighbourhood determinants.

Among 18 studies, 10 (L01, L03, L08, L09, L10, L11, L12, L14, L17, L18) included the examination of locality-specific determinants in relation to the observed TB pattern. Risk factors were identified through visual inspection of maps (L14), correlation analysis (L03, L09, L10, L17) and statistical modelling (L01, L08, L11, L12, L18). In the 10 studies reviewed, a total of 20 surrogate markers for 10 neighbourhood determinants were identified. These determinants were furthered categorized into four domains, namely, demographic, socioeconomic, environmental and care factors. As shown in Figure 2.3, a majority of neighbourhood determinants belonged to the environmental domain.

Figure 2.3 Number of neighbourhood determinants in each domain

Neighbourhood determinants identified from 10 studies were broadly categorized into four domains, namely, socioeconomic, environmental, demographic and care factors. A majority of the determinants belonged to the environmental domain, highlighting the relative importance of environment in influencing TB distribution



2.4.1 TB and demographics

Age and immigration status were two demographic determinants identified in the studies. Chan-Yeung (L08) found that the spatial units with higher percentage of elderly population had a higher TB notification rate in Hong Kong. Ageing is a well known factor attributing for TB disease progression, as host immunity will wane as age increases and dormant TB bacteria would be easily activated. As the demographic paradigm shifts in Hong Kong, it is generally assumed that TB in the elderly reflects endogenous reactivation of an earlier infection.

Immigrant status was found to be significantly associated with TB in 4 studies (L03, L12, L14, L17). Immigrants from regions of high TB prevalence were contributing

to the TB epidemiology of destination region via importation of latent infection at their arrival (L03, L14, L17), and through continued contacts to their country of origins (L12). Moreover, Davidow (L14) also showed that migrants in New York City were likely to receive less treatment than the resident population (L14). While other studies analysed the influence of international immigrants, Zhong (L03) analysed the internal migrant in China and identified a significant association with TB in Beijing.

2.4.2 TB and socioeconomic status

All surrogate markers in socioeconomic domain actually reflect the degree of poverty in general. Poverty is multidimensional in nature, the effect of which could be expressed by numerous surrogate markers. Six studies have identified a total of 5 neighbourhood determinants as a surrogate of poverty (Table 2.3). Determinants including income level, employment status and use of public assistance are regarded as direct measure of poverty, since these determinants directly reflect the actual income received. Determinants including education attainment and type of housing residing, which reflects the possibility of getting better paid job and the financial affordability respectively, are regarded as indirect measure of poverty.

Table 2.3 Risk factors indentified under different domains

A total of 20 surrogate markers for 10 determinants were recognized. As shown in the table, surrogate markers for a determinant could be different, implying that explanation in context is important in analysing the determinants and TB pattern.

Domain	Neighbourhood determinants	Surrogate marker
Social Economic	Economic condition	Housing type (L01)
		Education attainment (L08)
	Poverty	Income level (L08, L18)
		Employment status (L11, L12)
		Use of public assistance (L18)
Environmental	Overcrowding	House with TV and radio (L01)
		Badly maintained and unmodernised house (L17)
		Household crowding (L12, L17)
	Proximity to TB case	Household contact (L01)
		Distance from high incidence area (L09)
	Neighbourhood environment	Population density (L11)
		Temperature (L18)
		Number of shebeens (drinking place) (L11)
Demographic	Age	Elderly population (L08)
	Immigrants	Immigrants (L03, L12, L14)
		Ethnicity (L17)
Care factors	Treatment provision	Patient undergoing treatment (L01)
		Coverage of chest X Ray (L18)
	Treatment adherence	Patient lost to follow up (L01)
	Proximity to treatment	Distance to treatment centre (L01, L10)

Four studies (L08, L18, L11, L12) found that areas of higher TB rate were associated with a higher level of low income, unemployment and number of people receiving social assistance. In Munch's study (L11), TB cases and shebeens, a kind of drinking place in Africa, were clustered in areas of high unemployment. The correlation was significant when more than 40% of the population in an area was unemployed. In the same study, unemployment was also treated as indicator of behavioural pattern. This behavioural indicator were used to explain in a way that unemployed people might indicate more time spent in places of high transmission risk, e.g. shebeens, such that higher TB rate was resulted.

Two studies (L01, L08) adopted an indirect measure to illustrate the effect of poverty. In Chan-Yeung's study (L08), low education attainment was found to be a significant determinant of TB rate. Though the authors did not elaborate why low education attainment was treated as a surrogate of poverty, it could be easily understood that high education attainment is essential to secure a better paid job in a metropolitan city where tertiary and quaternary industries dominate.

2.4.3 TB and the environment

A total of 8 environmental determinants were identified. Higher TB rate was observed in areas of overcrowding, close proximity to infectious agent, and areas with unfavourable conditions.

Overcrowding has been historically a determinant of TB, as it increases the risk of personal contact and facilitates disease transmission if an infectious patient is

present. As transmission of TB requires prolonged exposure to an infectious agent in a confined area, household crowding is thus often a concern. Household crowding is usually quantified as the ratio between rooms and person in a household. In Kistenmann's study (L12), the number of rooms within flats per resident was used as a surrogate of household crowding.

Indirect surrogate markers for household crowding were also noted in this review. Randremanana (L01) hypothesized that watching television together increased the contact between household members, which reflected a certain degree of overcrowding and thus increased the risk of TB transmission. Therefore, ownership of televisions was used as an indicator of overcrowding, given that it was also an indicator of time spent at home. However, this assumption is only valid for study areas where social mixing is limited, so that majority of transmission of TB occurred within household. While it is evident that close proximity to household contact increases the risk of TB infection, Yeh (L09) demonstrated the proximity concept in the larger scale. In his study, dose-response relationship between high incidence mountainous areas and surrounding townships were detected. TB risk for surrounding townships decreased as the distance from mountainous area, where TB incidence was high, increased.

Six studies attempted to investigate possible relationship between TB and unfavourable neighbourhood environment, e.g., population density, and five of them failed to find statistically significant association with TB. Chan-Yeung (L08) stressed the need to find a better measurement of population overcrowding, as there

is currently no marker for evaluating the influence of overcrowding on TB incidence in Hong Kong.

2.4.4 TB and care factors

TB treatment is not only effective in curing TB disease but also effective in reducing disease transmission. Four determinants about provision of TB treatment were found significantly associated with TB pattern in 3 studies. As shown by the studies conducted by Yanagawa (L18) and Jacobson (L10), availability and accessibility of TB services, illustrated as coverage of chest X-ray and distance to treatment centre respectively, were important determinants to area-specific TB rate. Randremanana (L01) also highlighted the importance of treatment adherence as he found that areas with more patient stopping treatment generally had a higher TB risk.

2.5 Techniques applied in studying tuberculosis epidemiology

Techniques for constructing a spatial database, mapping disease and performing spatial analyses are discussed in this section.

2.5.1 Constructing spatial data

Spatial data forms the basic element of all spatial analyses. Spatial data in studies reviewed are classified in two types, namely, the spatial data of TB and spatial data of other factors. Getting spatial information for TB data is usually done by geocoding. Geocoding refers to the process of assigning a location, usually in the form of coordinate values (geocodes), to an address by comparing the descriptive location elements in the address to those present in the reference material. Geocoding requires patient's residential address, therefore it is not surprising that all TB data used in the studies were obtained from national surveillance system, where confidential data like residential address of patient were kept. Twelve studies reported the process of geocoding TB data using various geocoding engines and 1 study conducted in remote rural areas combined geocoding with use of GPS to obtain accurate coordinates (L10). The successful geocoding rate, ranging from 70%-100% among the reviewed studies, appeared to be satisfactory, given the high validity of address and availability of geocoding engine.

Geocoded data appears as a point, a one-dimensional structure. Point data are particularly useful for cluster detection (L11), outbreak detection (L04) and measuring the proximity between cases and other risk factors (L05). However,

ecological effect of neighbourhood determinants could hardly be measured at point-level, therefore, 16 out 18 studies aggregated data into various spatial units with only two studies (L07) merely focused on the point distribution of TB in relation to TB treatment (L07) and molecular clustering (L13).

Spatial unit can be defined as the geographic boundary which researchers were aggregating their data. Most of the studies adopted geographic boundary delineated by census (Table 2.4), generally due to three reasons: 1) census areas contain populations reasonably homogeneous with regard to socioeconomic composition; 2) availability of data for comparison, and 3) to protect spatial confidentiality. By using area-based measurement, various databases containing neighbourhood data, TB epidemiological data, demographic data and health services data, were linked. Investigation of contextual neighbourhood effects on TB epidemiology became possible. Number of spatial units implies the spatial resolution as well as the sample size of the study. In the 17 studies using aggregated spatial data, the number of spatial units ranged from 21 to 1835, from ward area in the UK to dwelling units in Africa.

None of the study discussed the statistical reliability in small area analysis. Choosing the right spatial units can be difficult. A study reported that inappropriate choice of spatial units may lead to inconsistency of results (L05). Some studies reported the modifiable area unit problem (MAUP). The MAUP arises because the value resulting from the aggregation can be different if the artificial polygon

boundaries are modified. This is particularly troublesome for study investigating spatial and temporal change of TB epidemiology.

Table 2.4 Spatial unit used in each study

Both point and area data were used as the analytical units in studies. A majority of them grouped TB data into area units and census tracts were the most commonly used spatial unit.

Code	Type of spatial units	Geometry of spatial unit	Number of spatial units
L02	Country	Area	53
L03	District	Area	18
L04	Census tract	Area	76
L05	Census tract	Area	539
L06	Census tract	Area	109
L08	Large street block group (LSBG)	Area	430
L09	Townships and surrounding areas	Area	>30
L12	Subdistrict	Area	78
L14	Health district	Area	30
L15	Dwelling unit	Area	1835
L16	County	Area	Unknown
L18	Prefecture	Area	Unknown
L01	Neighbourhood / Treatment centres	Area / Point	192 / 16
L10	Census tract /Treatment centres	Area / Point	Unknown / 5
L11	Enumerator areas / Shebeen	Area / Point	39 / 58
L17	Ward area / Case	Area / Point	21 / 444
L07	Treatment centres	Point	13
L13	Case	Point	182

2.5.2 Disease maps used

Disease map is an inevitable tool for visualization of disease pattern. Maps were not primarily used to provide geographic information about TB but also to facilitate investigation of relationship between observed pattern and hypothesized determinants. A total of 80 maps were found in the 18 studies, giving an average of 4.4 maps for one study. Based on the cartographic function, maps used are classified into 5 categories as shown in Table 2.5.

Table 2.5 Classification of maps based on cartographic function

Five types of maps serving different functions were identified from the reviewed studies.

Types of map	Definition
Choropleth map	Features are symbolized as shaded colours within a predefined boundary. The boundary of mapping units is usually the same of the spatial unit adopted in the study
Dot map	Specific location of features, eg, TB case and treatment centres, were mapped
Cluster map	Cluster of TB cases and neighbourhood determinants were mapped
Buffer map	Distance between cases and/or other neighbourhood determinants were mapped
Chart map	Values of features were displayed as a chart on the map

Since a map can be composed of different layers and a layer can be mapped for different functions, different functions can be served in one single map. Figure 2.4 shows the number of different map functions applied in each study. A majority of studies (n = 12) adhered to one mapping function only, that is, choropleth map. A maximum of 3 mapping functions were applied in a study.

Figure 2.4 Pie chart showing the maximum number of map functions of each study

All the maps in each study were examined. As a result most of the studies adhered to one mapping function only (i.e. choropleth mapping)

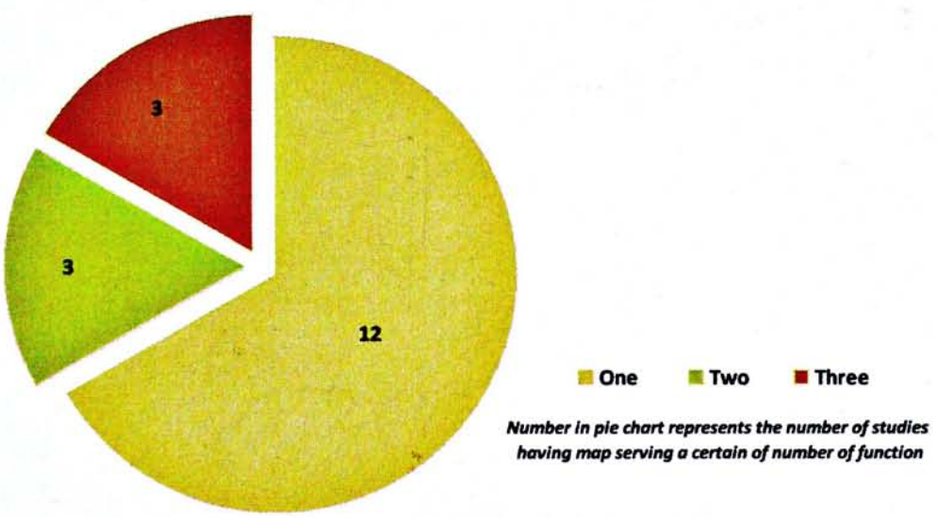
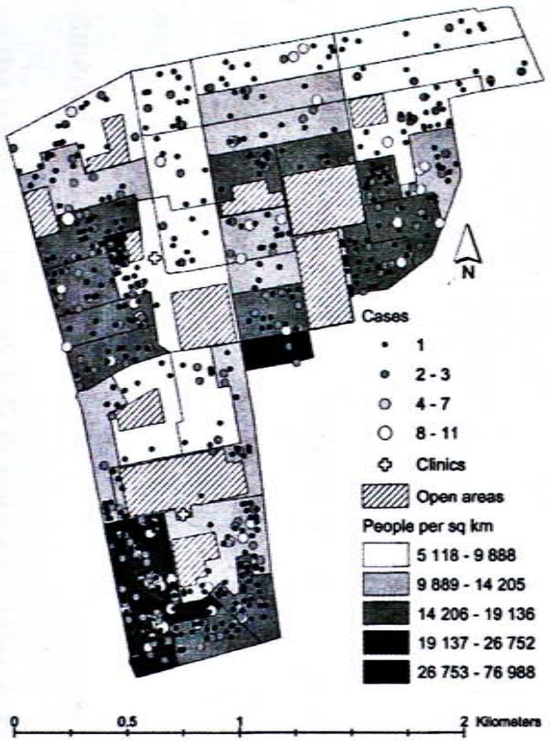


Figure 2.5 shows the maps extracted from Munch's study (L11) as an example to illustrate how three different mapping functions were used in his study. As shown in the figure, Munch used a single value dot map to represent the distribution of clinics, a multi-value dot map to represent various number of cases at specific locations and a choropleth map to illustrate the population density in each

enumerator areas. He then overlaid a cluster map showing different types of clusters in relation to the level of unemployment, which was shown as a choropleth map.

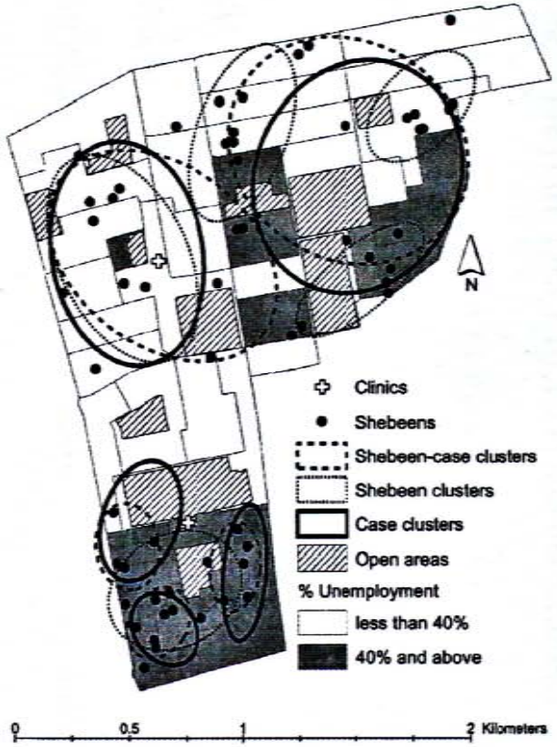
Figure 2.5 Maps used in Munch's study, as an example of different mapping functions

These are maps extracted from Munch's study (L11) to show how different functions of mapping were applied to illustrate various spatial phenomenon. In the study, Munch used a dot map, a multi-value dot map, a choropleth map and cluster map to show the distribution of clinics and shebeens, the distribution of TB cases, the population density and unemployment level and detected clusters respectively in Figure 2.5a and Figure 2.5b.



Source: Munch et al, 2003

a) Map showing the distribution of TB cases in relation to other neighbourhood determinants



Source: Munch et al, 2003

b) Map showing the geographic hotspot of TB cases

Mapping techniques are often restricted by the geometry of data, results from spatial analysis and the availability of software for mapping. Of the studies reviewed, choropleth map and dot map were the most popular types of map used. Dot map was used for mapping individual TB cases (L05, L08, L10, L11, L13, L17) locations of treatment centre (L01, L07, L10, L11) and the locations of high risk facilities (L11) .While it is acceptable for mapping the exact locations of treatment centres or other non human features, however, there is growing concern over spatial confidentiality when mapping exact locations of patients. Several studies successfully identified patients's address from published maps and they called for the awareness of breaching patient's confidentiality through maps. An additional problem of using dot map is the difficulties in deriving summary statistics of TB.

Choropleth mapping of TB case can be one of the means to protect spatial confidentiality. However, it is always not encouraged to map exact case count as the disease distribution is heavily influenced by the underlying population. Mapping absolute number of TB cases per area may be misleading.

As an alternative, a number of studies tried to map disease distribution in terms of relative risk (L01), excess risk (L02) and disease rate (L09, L12, L14, L15. L16, L18). Mapping statistical figures tends to provide a more reliable visualization of disease pattern, though interpretation should still be cautious. Mapping is not only applied for visualizing TB pattern. Researchers also mapped pattern of other risk factors such as percentage of elderly (L08) and distance to high incidence area (L09). By overlaying the pattern of TB with risk factors on a map, maps were

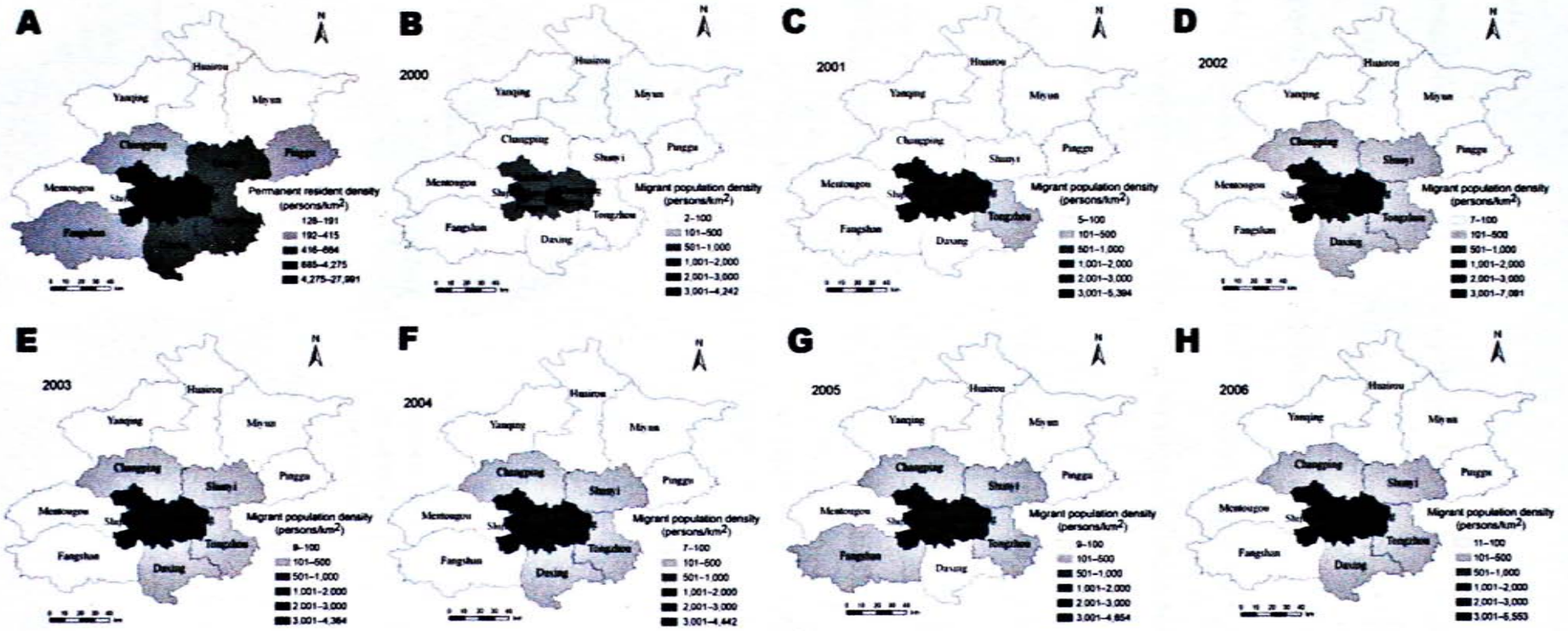
treated as a correlation analysis by visual inspection, though the correlation was not necessarily established statistically (L10, L14, L16).

The method used for data classification of map is also an area of concern. Inappropriate use of classification may lead to confusion and even false interpretation. Moreover, only a minority of studies mentioned about how the data was classified in maps while the majority of article did not justify nor explain the classification scheme they used for maps. In fact some studies adopted a series of maps for visualizing the temporal change of spatial distribution, which posed a greater challenge to data classification. In order not to cause confusion in interpretation, the classification scheme should be standardized for the purpose of comparison over time. Figure 2.6 shows a map extracted from Jia's study (L03). This figure serves a good example of appropriate use of data classification to visualize disease pattern over time.

Finally, locator maps were found in 6 studies. Locator maps represent the geographic location of the study area in a broader smaller scale, which can familiar readers with the study area and enhance the communicability of disease maps.

Figure 2.6 Maps used in Jia et al's study (L03), as an example to illustrate the classification intervals for a series of map over time

The data classification for mapping temporal change of spatial distribution should be standardized. As shown below, the classification interval used in all maps was the same which allows comparison among maps over different time.



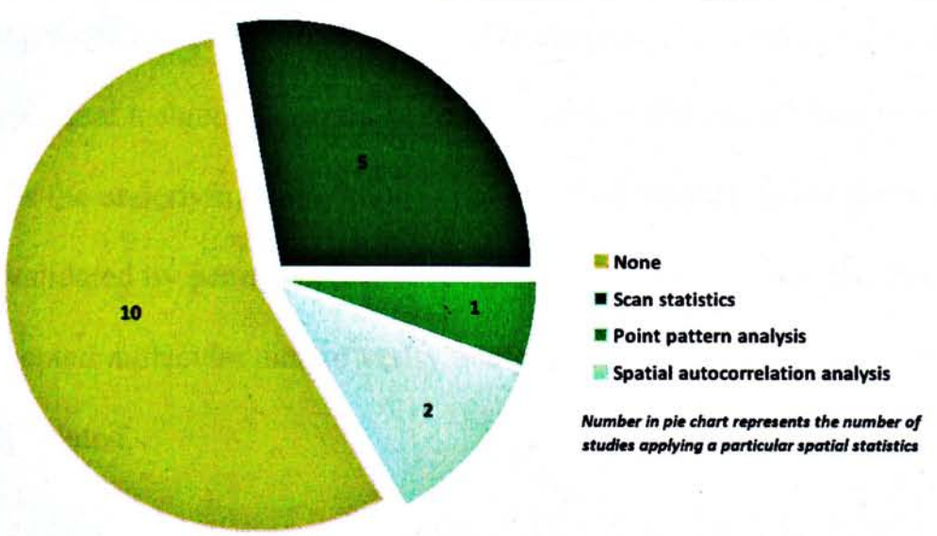
Source: Jia et al, 2008

2.5.3 Integrated approach using spatial statistics, conventional statistics and molecular analysis

In the 18 studies describing spatial pattern, 8 of them used spatial statistics to quantify the TB pattern. Moreover all the 8 studies using spatial statistics were published 2003 onwards, at the times when technological advance allows spatial statistics packages be installed and executed in GIS platforms. Figure 2.7 shows the spatial statistics used to detect clusters. According to the techniques involved as well as data types required, spatial statistics are grouped into three types, namely, point pattern analysis, spatial autocorrelation analysis and spatial scan statistics.

Figure 2.7 Pie chart showing the number of studies using spatial statistics

Serving different purposes, scan statistics, point pattern analysis and spatial autocorrelation analysis were used to identify spatial cluster for areal data, to identify spatial cluster for point data and to examine the spatial dependence among data respectively.



As a result of applying spatial statistics, 6 studies (L01 L02, L04, L05, L06, L11) were able to determine specific areas of significantly high incidence of TB, allowing further investigation to be performed. Among the articles using spatial statistics, a majority of studies applied scan statistics supported by SaTScan, a software developed to investigate disease cluster in terms of space and time.

The popularity of the scan statistics is due to three reasons: 1) SaTScan is an automated process to calculate a likelihood ratio and scan for areas that of significantly high likelihood ratio; 2) Underlying population size is considered in measuring TB clusters; 3) The software is free of charge. However, the operation of SaTScan involved parameter settings that are sensitive to user's choice and demands thorough understanding of the spatial distribution and relationship of the data. This in turns implies that SaTScan may not be suitable for exploratory studies. The use of spatial autocorrelation and point pattern analysis could detect spatial clustering based on the inherent spatial arrangement. The results evaluated from these analytical techniques would be more objective but careful interpretation is needed as the underlying population is not adjusted. Spatial cluster detected was further validated by genetic evidence. Four (L04, L05, L08, L13) studies supplemented molecular data to verify that genetically related cases were also spatially related.

Researchers further validated possible contribution of neighbourhood determinants to the clustered pattern using regression analysis. However, none of the studies adopted geographic regression and spatial modelling techniques to examine the

relationship between TB and its determinants. Instead, conventional statistics like Chi-square test, multilevel analysis, multivariate modelling were used. It should be noted that, spatial unit in conventional statistics are treated as independent entity, assuming that the spatial arrangement of these spatial units as well as its associated values do not pose any effect in the observed pattern. This assumption violates the basic assumption of spatial analysis that value in one place is influenced by values in proximate places. Yanagawa in his study stated an important observation from his regression model. In his study, he examined the spatial distribution of residuals resulted from his conventional regression models. He found that greater residuals were observed in spatial units located in the northern part of Japan and he suspected possible influence of place-factors in northern part was contributing to the greater residuals.

2.6 Research gap and thesis objectives

The literature cited collectively highlights the interrelatedness of geography and TB, in terms of disease transmission and progression. The varying pattern of TB is likely to be due to the underlying distribution of risk factors, in terms of biological, psychological, socioeconomically and environmental. By using a spatial approach, researchers were able, in summary, to: 1) describe the geographic distribution of TB; 2) identify clusters of significantly more TB; 3) assess the place-specific association between TB and neighbourhood determinants from various aspects and 4) predict epidemic and planning and targeting interventions.

2.6.1 Research gap

If the study of TB epidemiology and its neighbourhood determinants is to further etiological understanding and facilitate control strategies, it is crucial to have effective models and theories on how neighbourhood determinants at different scale may influence TB distribution. What is not yet understood is the spatial relationship of the association between TB and its determinants. Studies reviewed successfully identified significant determinants of TB and these determinants were assumed to be posing equal effect to the whole study area. Poverty is a well known determinant to TB, however the effect of poverty concentration has not been rigorously examined. In this sense, despite researchers were able to establish relationship between TB and determinants in absolute term, they failed to consider the possible influence on TB epidemiology if these determinants are spatially clustered. To address the spatial effect, research is needed 1) in the use of GIS technology to both

visually identify and empirically measure spatial relationships of geographic, environment and social influences on TB and; 2) to apply methodologies for contextual investigation of the spatial interdependence and spatial dynamics.

2.6.2 Thesis objective

In conclusion, this review raises several important questions that warrant exploration through further research. In order to fill the remained research gap, the thesis is set to

1) describe the TB epidemiology in Hong Kong using a new spatial unit
2) apply spatial statistics to quantify the spatial pattern and hotspots of TB
3) account for the locality specific determinants to TB using a new spatial modelling technique called geographically weighted regression, which accounts for the spatial non-stationarity in the association.

CHAPTER THREE METHODOLOGY

3.1 Rationale and approach

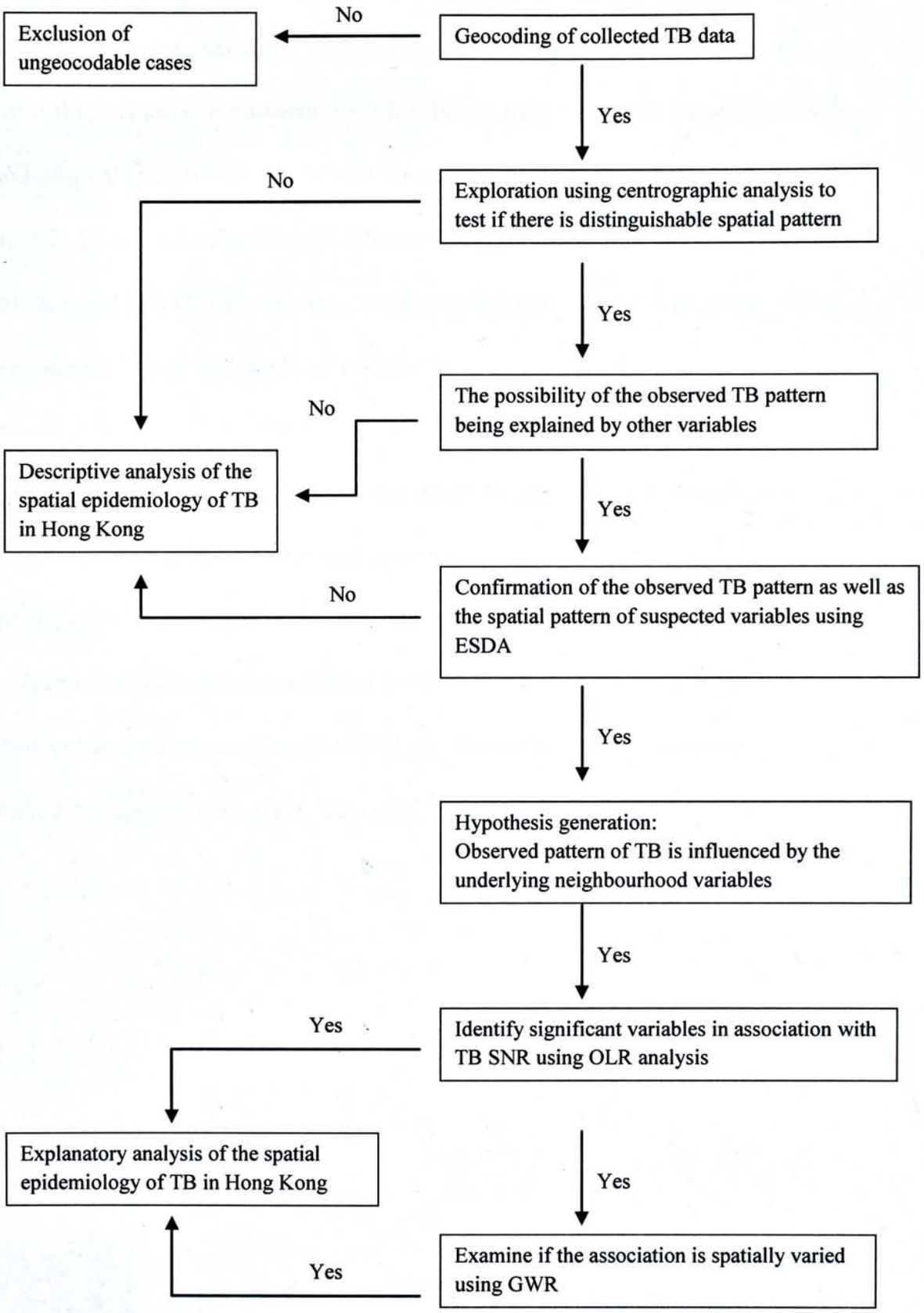
The importance of spatial perspective in studying TB has been well documented. In this study, taking Hong Kong as an example, TB epidemiology and its associated neighbourhood factors were investigated using a combination of spatial analytical techniques.

3.1.1 Logical flow of the study

The whole study started from the construction of TB spatial database (Figure 3.1). Collected TB data would undergo a process called geocoding. As only data with spatial reference could be analysed by spatial analytical techniques, ungeocodable TB data would therefore not be valid for further analysis. Data could not be geocodable because of various reasons, for instance, the validity of reported

address or the capability of the geocoding engine. Centrographic analysis, which provides basic descriptors for a spatial pattern, of geocoded cases would be performed for initial exploration. If specific spatial pattern was observed, suspicion over possible neighbourhood factors in affecting observed TB pattern became reasonable. After reviewing and collecting possible and plausible neighbourhood variables that might explain the observed TB pattern, an Exploratory Spatial Data Analysis (ESDA) could be conducted. ESDA could help to confirm if the observed pattern as well as the variables demonstrate statistically significant spatial pattern (Lai & So & Chan, 2009). If clustering effect of TB was confirmed, any clustering of suspected variables might well be able to explain the observed pattern. Therefore, for those variables demonstrating significant clustering effect deduced from ESDA would be put in an Ordinary linear Regression (OLR) model to examine its association with TB. Significant variables would be further analysed in a Geographically Weighted Regression (GWR) model to examine if the association between TB and variables were spatially different.

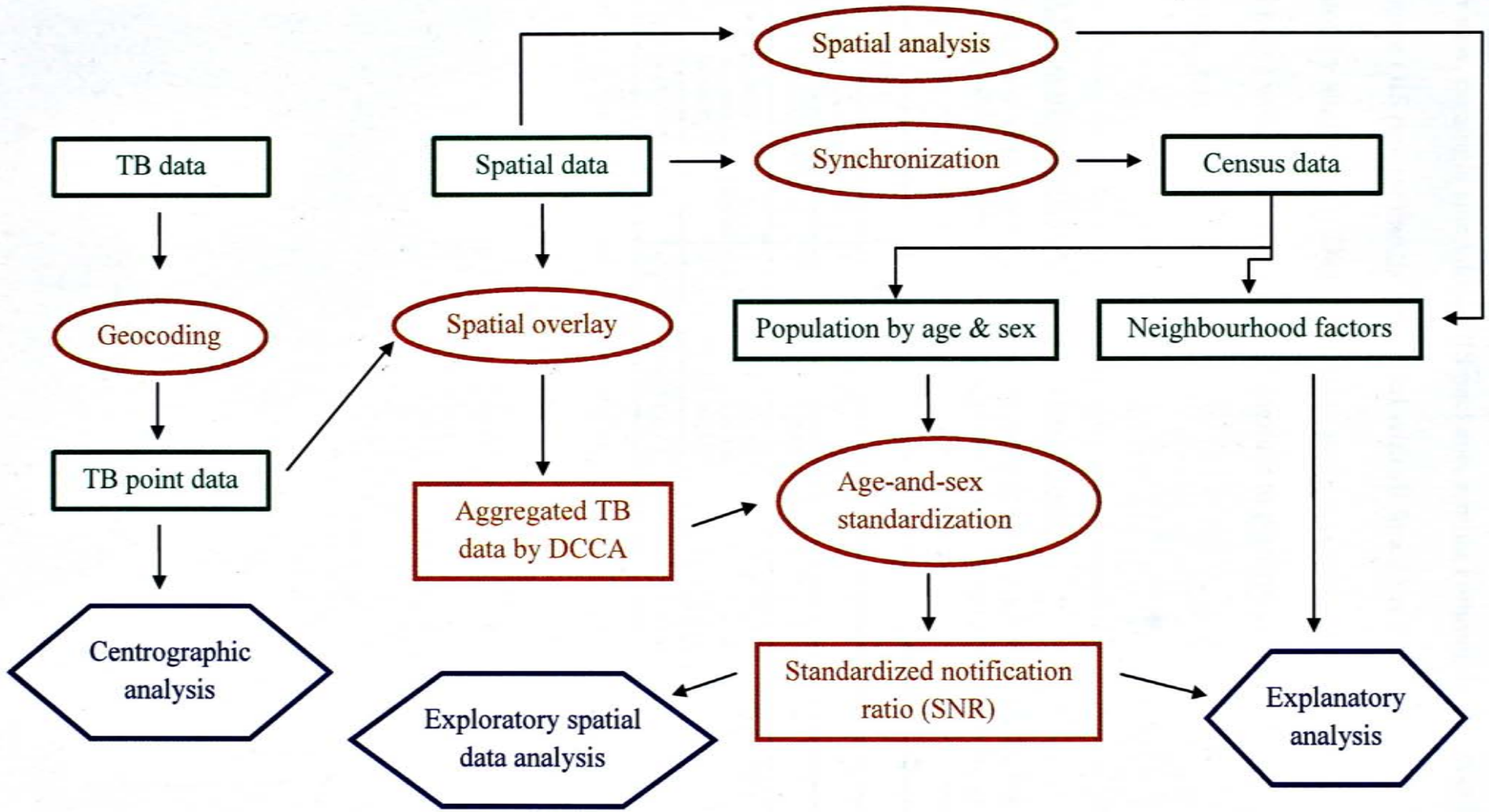
Figure 3.1 Schematic diagram showing the logical flow of the study



3.1.2 Methodological flow of the study

As spatial epidemiology research requires data from various sources, this study was therefore carried out within the environment of a geographic information system (GIS), which is a powerful tool for data management and integration. The methodological framework was briefly categorized into three stages as shown in Figure 3.2. The initial stage included the collection of raw data including TB data, spatial data and neighbourhood data, as shown as items coloured in green. These datasets would be manipulated and standardized to a common format in the second stage (items coloured as orange) so that all datasets could communicate with each other and analytical procedures could be performed. Manipulation processes included geocoding, spatial overlay, a series of spatial analysis, synchronization and standardization. Constituting the last stage of the study were three types of analysis (items coloured in blue). Centrographic analysis, exploratory spatial data analysis (ESDA) and explanatory analysis were conducted for data visualization, data exploration and data modelling respectively.

Figure 3.2: Schematic diagram showing the methodological framework of the study



Shape: Rectangle = Datasets; Oval = Process of data manipulation; Diamond = Analysis,
Colour: Green = First stage of study; Orange = Second stage of study; Blue = Last stage of study

Since the application and operation of GIS, particularly in the field of epidemiology, is rather new, no single standalone GIS package, neither commercial product nor open source GIS programme, is embedded with all functions that a spatial epidemiology study needs. Therefore in this study, a combination of statistical and GIS packages were used simultaneously in order to perform different analytical tasks (Table 3.1).

Table 3.1 List of statistical and GIS packages used

Since there is no single standalone GIS programme embedded with all the functionalities needed in this study, a variety of GIS programme were adopted for specific analysis.

Programme	Function
CrimeStat Version 3.0	Centrographic analysis
GeoDa Version 0.9.5.i	Exploratory Spatial Data analysis
SPSS Version 13	Explanatory analysis: Conventional regression analysis
GWR Version 3.0	Explanatory analysis: Geographically weighted regression
ArcGIS Version 9.2	Spatial analysis, Cartography (Visualization)

3.2 Choosing spatial units

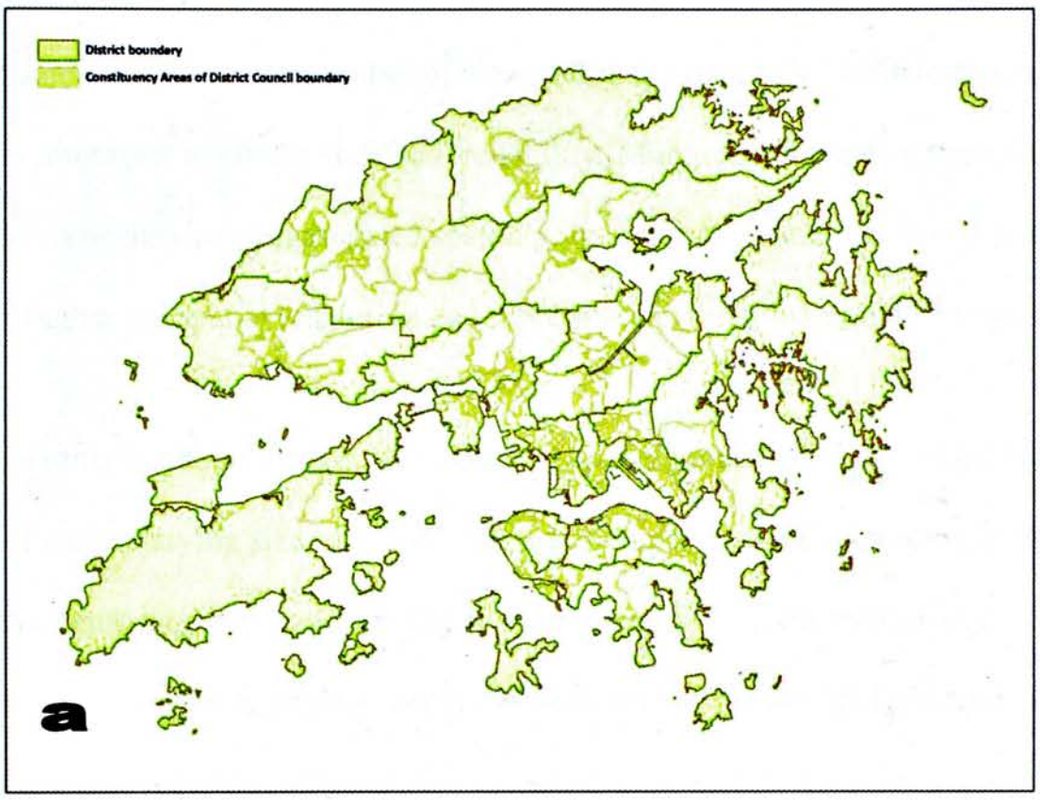
Adopting an ecological approach, data in spatial epidemiology studies are usually aggregated into and analysed by a particular spatial unit. However, choosing an appropriate spatial unit is always a technical challenge. Results of analyses, despite using the same dataset, could fluctuate greatly when different spatial units were applied. Since there is no universal standard for choosing an appropriate spatial unit, some criterion noted in literature review were considered as the selection criteria for spatial units in this study. In this section, a preliminary review of spatial units in Hong Kong was conducted to select a spatial unit for this study.

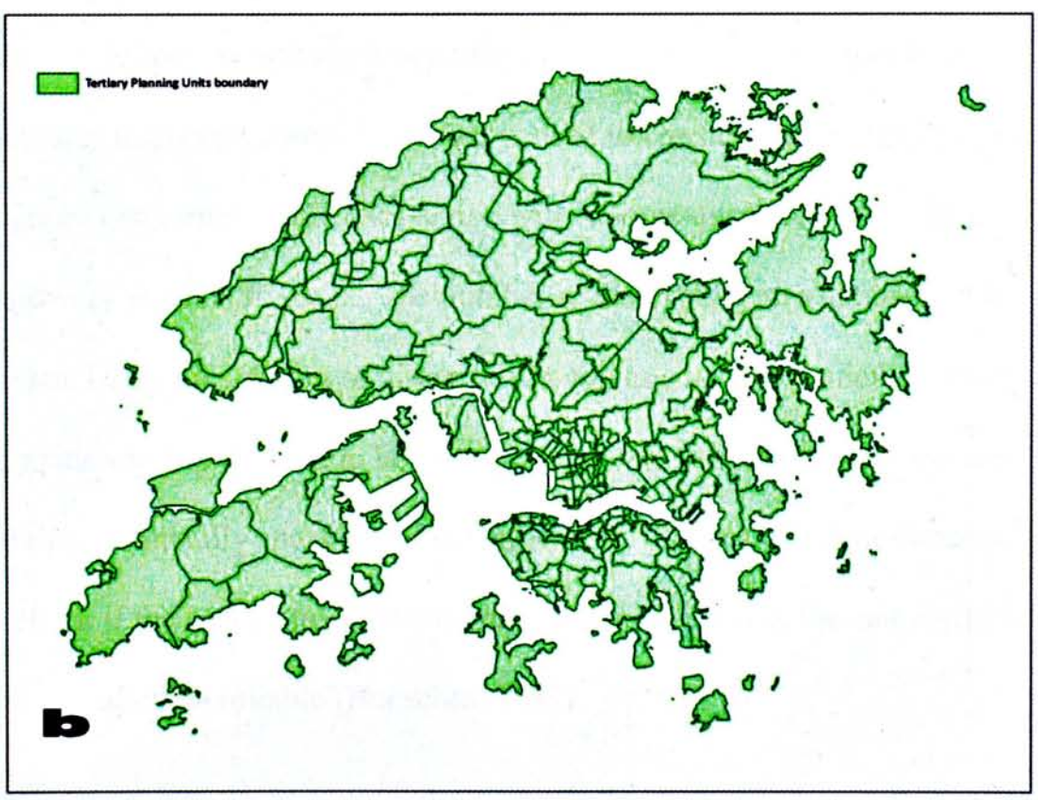
Three spatial units in Hong Kong, namely District Council (DC), Tertiary Planning Unit (TPU) and District Council Constituency Area (DCCA) were considered for potential spatial unit to be adopted in this study. DC, TPU and DCCA are in different scale and serve different purposes. DC is commonly used by government departments for administration and Census and Statistics Department to summarize population data, while DCCA is delineated according to a certain population threshold by Electoral Affair Commission for DC election every four years. As shown in Figure 3.3a, DC and DCCA belong to the same delineation hierarchy, in which case Hong Kong was divided into 400 DCCAs under 18 DCs in 2003 and each DCCA will return one DC member.

TPU is under a different delineation system from DC and DCCA (Figure 3.3b). It is demarcated by the Planning Department for town planning purposes. TPU is the third level of planning units under the hierarchy of planning units in Hong Kong. It is delineated when boundaries can be easily referenced by an adjoining feature, like boundary of a major development site and road centreline. In 2003, a total of 197 small TPUs were delineated.

Figure 3.3 Map showing the boundary of (a) DC, DCCA and (b) TPU

Hong Kong has 18 districts, which are further subdivided into numerous DCCAs for election purpose. TPU is under a different delineation system as it is demarcated by the Planning Department for town planning purposes.





Desirable characteristics for spatial unit noted in the literature review were grouped into 5 categories in this section. Indicators for the corresponding characteristics of DC, DCCA and TPU are summarized in Table 3.2. Number of spatial units reflects two concerns of the study, i.e. sample size and spatial resolution. The number of spatial units equals to the number of observations in the study. Sufficient sample size is important to ensure statistical reliability. Moreover, number of spatial units implies how data were aggregated spatially. In general, greater number of spatial units, higher the spatial resolution and more spatial variability could be revealed.

On the other hand, an increase in number of spatial units means a decrease in the size of each. Varying size of spatial unit may hamper pattern recognition in map because units larger in size often catch more attention than the smaller ones. Ensuring homogeneity of area size is essential to limit readers' natural tendency to associate area size with magnitude. As spatial units become smaller in size, number of TB cases enclosed in a particular unit will decrease. This concern is particularly important when case count is used to construct estimates of disease incidence. The estimates of disease risk will become unstable for spatial units that contain very small case count. The number of underlying population is also important in the calculation of disease estimate, as smaller population size of a geographic unit would tend to have an extreme rate because of the variance instability. Generally speaking, if the population number (the denominator) is less than 30, or if the case number (the numerator) is less than 5, the calculated rate would become less reliable (Buescher, 1997).

Also noted in many studies, the availability of data at same level is also a major concern of researchers when choosing a spatial unit, particularly for studies that focus on exploring association between TB and risk factors.

As some of them are theoretically contradictory, it is unlikely to be possible to have a spatial unit possessing all desirable characteristics. From Table 3.2, DCCA appears to possess most desirable characteristics by having the greatest number of spatial units, the smallest variation of area size and population size, moderate number of data available at the same level, and reasonable number of minimum cases count (more than 5 cases) per unit. Therefore in this study, a total of 400 DCCAs would be taken as the study unit and data will be aggregated to this level for analyses.

Table 3.2 Characteristics of three spatial units in Hong Kong

Different indicators were used to represent desirable characteristics of spatial units in Hong Kong. As shown in the table, DCCA possesses more desirable characteristics than the other two units.

Desirable characteristics	Indicators	DC	TPU	DCCA
Sufficient number of spatial units	Number of units	18	197	400
Sufficient number of case count	Minimum number of TB cases	246	1	6
Population homogeneity	Mean of population number	381182	33634	17153
	Standard deviation of population	148160	40867	4454
Area homogeneity	Mean area size (km^2)	60.92	5.68	2.74
	Standard deviation of area size (km^2)	59.65	10.26	9.98
Availability of variables at same level	Number of census variables available	34	14	22

3.3 Data collection

Like any other analysis, the availability and quality of data is always an important concern. For spatial analyses, the requirement of data is more stringent in the sense that confidential data such as residential address is needed. In Hong Kong, the TB dataset is relatively complete since virtually all known TB disease cases are reported and recorded. As no spatial sampling is needed, there is potentially no sampling error and bias in this study. Information of each TB patient, including one's residential address, is regularly collected at the time of notification. With the residential address, we are able to geocode each case down to building level. The high incidence of TB cases in Hong Kong provides us with a large sample size, meaning that there is little likelihood of encountering the problem of small area statistics. For spatial data, Survey and Mapping Office (SMO) of Lands Department has been maintaining a computerized land information system for mapping data and land boundary records covering the whole territory of Hong Kong. The digital spatial database at different scales is updated from time to time, allowing data of high spatial and temporal resolution available to the public. Moreover, the availability of up-to-date information on demographic and socio-economic characteristics of the population provided by Census and Statistics Department, together with TB notification system and Land information system, provides a favourable platform for pursuing spatial analyses of TB in Hong Kong.

3.3.1 Tuberculosis data

TB is a statutory notifiable disease in Hong Kong. As required by the Prevention and Control of Disease Ordinance (Cap. 599), any diagnosis of TB case should be notified to the Director of Health. In this study, notified TB cases from 2005 to 2007 were extracted from the surveillance system maintained by the Tuberculosis and Chest Service, Department of Health. Data including sex, presenting age, date of notification and residential address were used for analysis. Institutional approval was obtained from the Department of Health for access to the retrospective anonymised TB data in compliance with the Personal Data.

3.3.2 Spatial data

Spatial data used in this study included base maps and digital geographic features in Hong Kong. Base maps delineating the boundary of DC and DCCA published in 2006 were obtained from the Census and Statistics Department. Digital geographic features (Version 2.01) including building polygons and landscape features in scale 1:5000 were obtained from the SMO of Lands Department. All spatial data were geometrical in nature and projected to the HK Grid 1980 coordinate system.

3.3.3 Neighbourhood data

Neighbourhood data used in this study were categorized into census variables and spatial variables. Census variables were extracted from 2006 Population By-census report while spatial variables were generated with spatial data. The choice

of these variables was based on the literature review. Since DCCA was the spatial unit for the analysis, all neighbourhood variables were aggregated to this level. Selected variables were grouped into five domains to reflect key dimensions affecting the underlying variation of TB rate (Table 3.3). Considering the mountainous landscape of Hong Kong, some proportion of land area is barely inhabitable. The measurement of population density including uninhabitable area may not reflect the magnitude of population density in reality. Therefore in this study, land over 200 meters in altitude were considered uninhabitable and excluded in the calculation of population density. Only inhabitable area, where land area is below 200 meters, were used to calculate the population density in each DCCA. Building coverage was introduced in this study as a surrogate to reflect the degree of neighbourhood crowding. It referred to the percentage of area covered by buildings, which could be residential building, commercial building and housing structure, per DCCA.

Table 3.3 List of neighbourhood variables (n = 400)

A total of 14 neighbourhood variables were grouped into 5 different domains. Most of the variables were extracted from census report and two variables were generated in GIS.

Domains	Neighbourhood variables	Definition (in percentage)	Source
Immigrant Status	Non Hong Kong born	Population born outside Hong Kong	Census
	Short duration of residence	Population residing in Hong Kong less than 1 year	Census
Economic	Economically inactive population	Population including home-make, student retired person and other economically inactive person	Census
	Low income individual	Population having a monthly income less than HKD 6000	Census
	Population engaging in secondary sector	Population engaging in manufacturing and construction sector	Census
	Low income household	Household having monthly income less than HKD8000	Census
Family structure	Not married	Population being single	Census
	Household of small family structure	Household of one person or one unextended nuclear family	Census
Living condition	Large household Size	Household with more than 4 residents	Census
	Room shared by person	Average number of rooms per person	Census
	Crowded quarter[1]	Quarter with more than 4 occupants	Census
	Population living in public housing	Population living in public housing	Census
Neighbourhood Environment	Population Density	Number of population per unit of inhabitable area (km^2)[2]	Spatial
	Building coverage	Area coverage of building in DCCA [3]	Spatial

[1] Quarters refer to unit of accommodation as flats, houses and structures which could be used for the purpose of accommodation

[2] Inhabitable area refers to land area below 200 metres in altitude. This variable is expressed as km^2

[3] Area of building as a percentage of total area of DCCA

3.4 Data manipulation

Since collected datasets were presented as both individual-level (TB data) as well as aggregated-level (Neighbourhood data), therefore they would undergo a variety of (geo)processing procedures so that they could communicate with each other within the GIS environment.

3.4.1 Tuberculosis data

According to the validated residential address, the three-year TB cases were geocoded to building level, or estate level if specific building could not be identified. The coordinate system used in this study was HK 1980 Grid, which is a planar transformation from UTM 1984. Since there is no automatic geocoding engine in Hong Kong, geocoding was performed manually. Geocoded cases were subsequently computed and stored as point data in GIS programmes for visualization and centrographic analysis.

Geocoded TB cases were also aggregated into DCCAs by using spatial overlay in ArcGIS. As TB incidence rate is highly dependent on the age-and-sex composition of underlying population, adjustment for the population structure is necessary before the computation of a TB estimate. Crude notification rate of TB (number of TB case per 100, 000 population) was age-and-sex standardized using an indirect approach. The resultant TB standardized notification ratio (TB SNR) would be used as the dependent variable for both exploratory and explanatory analysis in

subsequent sections. Details of indirect age-and-sex standardization were provided in Appendix 1.

3.4.2 Spatial data

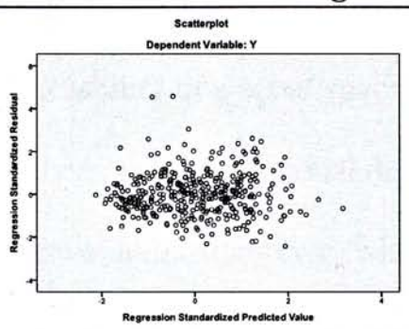
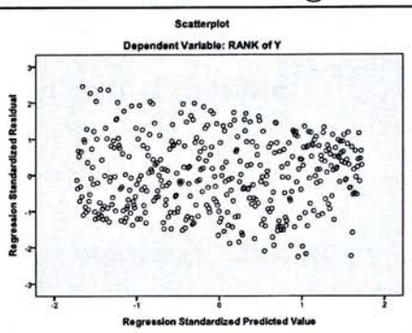
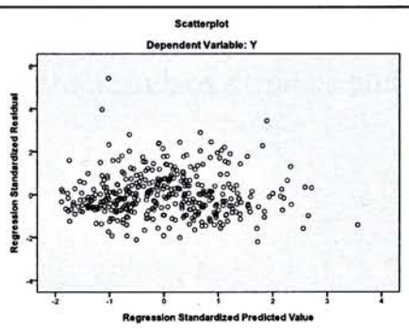
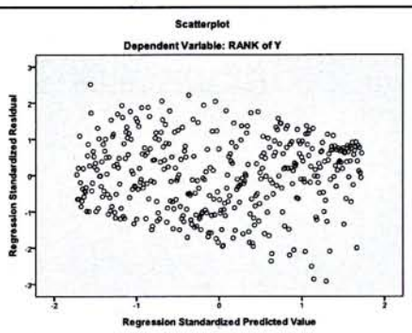
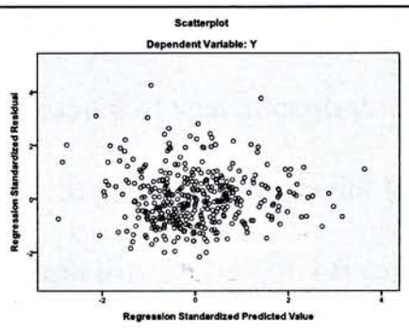
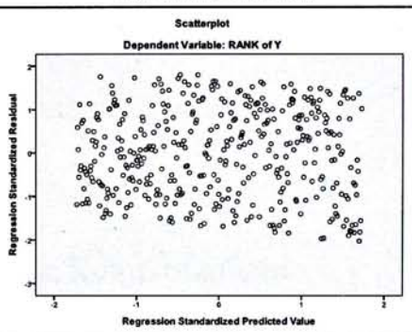
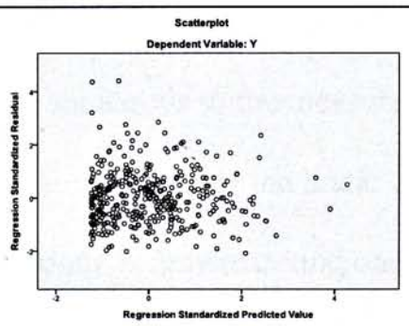
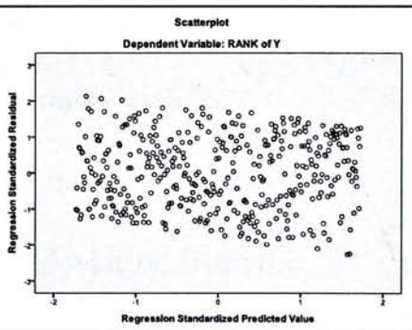
A relational geodatabase was built to store all spatial data in the file format called shapefile (.shp). Shapefile is a file format developed by ESRI, a GIS company, and could be read in most GIS packages. To georeference neighbourhood data from census report, coordinates of DCCA were attached to the dataset storing neighbourhood variables.

3.4.3 Neighbourhood data

All neighbourhood variables, except for the *population density* and *room shared by person* were expressed in percentage. To account for the problem of heteroscedasticity which usually exists in census data, a preliminary exploration using 4 selected neighbourhood variables were conducted. As shown in Figure 3.4, separate linear regressions were conducted and residuals were plotted respectively. The residual plots for un-ranked data showed some degree of clustering while the residual plots for ranked data appear to be fairly homoscedastic. Therefore in order to avoid the problem of departure from data normality, all neighbourhood variables, together with TB SNR, were ranked and ranked data would then be used for further analysis.

Figure 3.4 Residuals distribution of selected variables before and after data ranking

The residual plots for data before ranking showed certain degree of clustering while the residual plots for ranked data appeared to be fairly homoscedastic.

Variables	Before ranking	After ranking
Non Hong Kong born		
Low income household		
Not married		
Building coverage		

3.5 Centrographic analysis

Centrographic analysis, one of the descriptive measures for spatial data, provides the most basic type of descriptors of a spatial pattern. It incorporates spatial dimension into conventional statistics and provides summaries about the central tendency, dispersion and variability of a set of spatial data. The main difference between centrographic analysis and conventional descriptive analysis is that the former considers values in two dimensions (x,y) while the latter considers value in one dimension only. In this study, using TB point data, centrographic statistics including spatial mean centre, standard distance and deviational ellipse for TB distribution will be investigated.

3.5.1 Types of centrographic statistics

Spatial mean centre is a measure of spatial central tendency analogous to the classical statistics of mean. Represented as a point location, spatial mean centre was used to summarize the overall distribution of TB cases in Hong Kong. Standard distance was used to illustrate the degree of spatial dispersion and represented as a two dimensional data. It is analogous to the measures of dispersion such as variance in classic statistics. It calculates the linear distance from each point to the mean centre point, then a circle is drawn around one standard deviation from the centre point. Standard deviational ellipse is regarded as an extension of standard distance as it considers also the anisotropy of a point data set. An ellipse would be generated to illustrate the spatial dispersion, with major and minor axes reflecting the directional variation of the point pattern. A convex hull was also used to

illustrate the extent of spatial distribution which is analogous to the range in conventional statistics. Convex hull is a boundary drawn that circumscribes all the points in the distribution such that no point lies outside.

3.6 Exploratory spatial data analysis

Visual inspection of disease maps and centrographic statistics are useful for exploring possible spatial relationship, however individuals often have difficulty in recognizing random patterns or identifying specific clusters. Objective measures of spatial pattern are therefore necessary and quantitative metrics could be obtained by using ESDA.

The rationale of exploratory measures for spatial data is fundamentally built upon Tobler's first law of geography, emphasizing that everything is spatially autocorrelated. Using ESDA, it is possible to determine whether TB distribution in Hong Kong exhibits a significant deviation from an expected pattern. If the distribution is not random, then it becomes important to measure the magnitude and the nature (positive or negative) of the spatial autocorrelation in order to make relevant hypotheses on observed pattern and its underlying causes. Moran's Index and Local Indicator of Spatial Association (LISA), two commonly used spatial statistical instruments, would be applied in this study for the detection of global and local spatial autocorrelation respectively (Wong & Lee, 2009).

3.6.1 Spatial proximity matrix

Conducting analysis of spatial autocorrelation requires the enumeration of spatial relationship among spatial units. Techniques to enumerate spatial relationship vary according to the geometry of data. For areal data, two types of neighbourhood contiguity relationship can be created, namely, rook and queen. Rook contiguity

uses shared boundaries to define neighbours, meaning that neighbours are defined if they share a common boundary. The queen contiguity determines neighbours as those having common boundaries or common corners. Therefore the number of neighbours defined using the queen criterion should be the same or greater than that defined using rook criterion. In order to be conservative in this exploratory study, analyses in this study involving the measures of polygon adjacencies would adhere to the rook method. The enumeration of spatial relationship would be performed in GeoDa and results were expressed by a spatial proximity matrix, or sometimes called spatial weights file.

3.6.2 Moran's Index

Moran's Index is a classic indicator of global spatial autocorrelation. It signifies whether clustering of values exists and gives a summary value for all observations in Hong Kong as a whole. Moran's Index is interpreted very much like a correlation coefficient. The range of possible value is +1 to -1. Index near +1 indicates a strong spatial pattern, meaning similar values are located near one and other. Index near -1 indicates a strong negative spatial pattern, meaning dissimilar values are located near one and other. Index near zero indicates the absence of spatial autocorrelation, meaning the distribution of value is random.

3.6.3 Local Indicator of Spatial Association

While Moran's Index, as a global measure, tells us if there exits spatial clustering in the distribution of values, LISA, as a local measure, highlights specific areas

where the cluster of values are found. LISA is a local version of Moran's Index, as it decomposes the global measure into its contribution for each spatial unit and then detects similarities or dissimilarities in values of a spatial unit in relation to its neighbourhood. Instead of a summarized value, the results of LISA were generated for all DCCAs and results were visualized in the form of significance and cluster maps.

3.7 Explanatory analysis

Univariate analysis was performed to examine the relationship between TB SNR and all neighbourhood variables. To address the multicollinearity problem among census variables, separate linear regressions for each of the five domains were conducted to select neighbourhood variables that were least affected by the problem of multicollinearity. Multivariate analysis using OLR was performed to examine the association between TB and selected neighbourhood variables.

Linear regression gives a summarized (global) estimation of parameters for significant variables. According to the assumption of independence of errors and error variance homogeneity, these parameters are assumed to pose a constant effect to the dependent variable over the entire study area. However, if the values of TB SNR and neighbourhood variables are spatially autocorrelated, linear regression modelling may not be appropriate for analysing such a dataset because the assumption of data independence is violated. As a result, global estimates from the linear regression model will misrepresent the real phenomenon by ignoring the spatial relationship. If the relationships between variables do spatially vary, this neglected dimension in the OLR model will likely be confined to the error term (residuals). Moreover, if the spatial variation of relationship between TB SNR and neighbourhood variables does exist, the failure in including this variation can reduce the explanatory power of the OLR model. GWR thus aims at incorporating the measurement of spatially-varying relationships among variables so that spatial relationship is no longer treated as a residual. In this study a GWR model was calibrated to address the concern over spatial variability of variables and to

determine if the spatial non-stationarity of the relationship exists. Details on how GWR works is provided in Appendix 2.

3.7.1 Selecting variables for modelling

Spearman's correlation coefficient was calculated to determine the relationship between TB SNR and neighbourhood variables. Variables with a $P < 0.05$ were considered significant. Separate linear regressions using Enter approach was conducted for the five domains of neighbourhood data. Variables with the highest partial correlation in each domain were selected as independent variables in the OLR model for further multivariate analysis.

3.7.2 Ordinary linear regression

Multivariate analysis using OLR was performed to identify neighbourhood variables that were significantly associated with TB SNR. TB SNR, as a dependent variable, and selected neighbourhood variables in the previous regression analysis, as independent variables, were put into a multiple regression model. A backward stepwise conditional approach was adopted, with F probability to retain being <0.05 and F probability to remove being >0.10. Statistical analyses were performed using SPSS Version 13.

The OLS equation can be expressed as:

$$TB\ SNR = \beta 0 + \beta aXa + \beta bXb + \beta cXc + \beta dXd + \beta eXe + \xi$$

where β0 is the intercept term, βa, βb, βc, βd, βe are spatially varying coefficients of variables in five domains respectively, and ε is an error term, (u,v) are the coordinates of its position

3.7.3 Geographically weighted regression

The local regression model for this study was calibrated using GWR software developed by Fotheringham & Brunsdon & Charlton (2003). Four hundred centroids of DCCA were generated in GIS and taken as reference points for measuring distance between observations as well as regression points in the GWR model. A Gaussian GWR model was calibrated, using TB SNR as the dependent variable, and significant neighbourhood variables identified in the OLR as independent variables. The spatial weighting of DCCA was defined by an adaptive bi-square spatial kernel and optimum bandwidth for the kernel was chosen by minimizing the Akaike Information Criterion (AIC). Details of how AIC works is provided in Appendix 3.

The GWR regression model for TB SNR can be written in this format:

$$TB\ SNR\ (u, v) = \beta 0(u, v) + \beta a(u, v)Xa + \beta b(u, v)Xb + \beta c(u, v)Xc + \beta d(u, v)Xd + \beta e(u, v)Xe + \varepsilon\ (u, v)$$

where β0 is the intercept term, βa, βb, βc, βd, βe are spatially varying coefficients of variables in Domain a, Domain b, Domain c, Domain d, and Domain e attributes respectively, ε is an error term, and (u,v) are the coordinates of its position

As GWR calibrated separate regression equation in each DCCA, DCCA-specific results including parameter estimates, t-value, and R-square were computed for each DCCA. Each set of DCCA-based parameter estimates could be mapped, which helped uncover spatial variations in the relationship between the variables that would be unnoticed in the OLR analysis.

Whether GWR offers an improvement over the OLR could be determined through the comparison of AIC and residuals. Model of lower AIC is preferred. Moreover, ANOVA test was employed to test if the change in residuals was statistically significant, with a null hypothesis that the GWR model represents no improvement over the OLR model. Monte Carlo test was conducted to determine the significance of the spatial variability in the GWR parameter estimates, with the null hypothesis of random distribution of parameters estimates in each DCCA over Hong Kong. Details about Monte Carlo test were summarized in Appendix 4.

CHAPTER FOUR RESULTS

4.1 Overview

In this chapter, results are presented by following the methodological sequence outlined in Chapter Three.

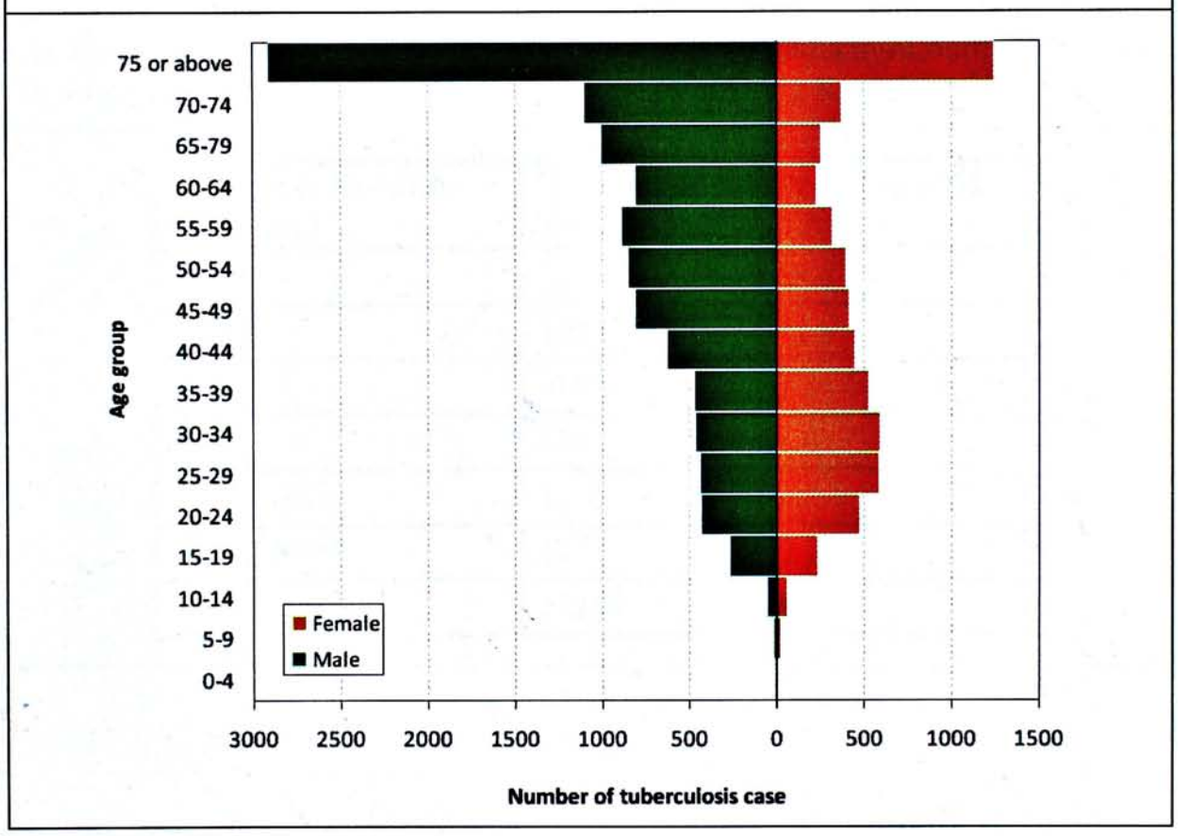
4.1.1 Individual level

A total number of 17389 new TB cases were notified in 2005 to 2007, of which 17294 (99.5%) cases were geocoded. Cases not geocoded were due to the following reasons: 1) Reported address was outside Hong Kong territory; 2) Reported address was incomplete or invalid and 3) No address was provided. Only geocoded TB cases were included in the study. Figure 4.1 shows the age and sex distribution of geocoded TB cases. The mean age of TB cases was 56 years old (SD

= 21.25), with a range from 2 years old to 108 years old. Forty percent (n = 6879) of cases were aged 65 or above, which were classified as old age TB in this study. In general, 64% (n = 11149) of cases were male, giving a male-to-female ratio of 1.81:1. Sex ratio varied in different age groups. A slightly more female cases than male cases was observed in younger age groups but the situation was reversed when age increased.

Figure 4.1: Population pyramid showing the number of TB case per age group

More TB cases were found in male and number of notification increased as age increased.



Based on the geocoded point location of TB cases, a distance matrix was computed to summarize the minimum distance between one case and all other cases (Table 4.1). The minimum distance between two cases was 19 metres on average, with a range from 0 metre up to 6437 metres. Ninety-one percent (n = 15802) of cases were residing within 50 metres of another TB case. As demonstrated by the minimum distance of 0 in Table 4.1, 69% (n = 11899) of cases were co-located with at least one TB case in the same building.

Table 4.1 Table showing the minimum distance between one case to all other cases

Minimum distance refers to the distance between one case and its nearest case. As shown in the table, about 70% of cases demonstrated a minimum distance of 0, implying the degree of co-locating TB cases in Hong Kong.

Minimum Distance (in metre)	Number of TB case	Percentage of TB case
0	11899	68.8
1 - 50	3903	22.6
51 - 100	916	5.3
101 - 500	513	3.0
501 - 1000	41	0.2
Over 1000	22	0.1
Total	17294	100

Table 4.2 provids further information on co-location of TB cases. All TB cases were found in 8355 unique buildings over Hong Kong, 31% (n = 5395) of them were located in buildings that did not have any other cases while the remaining co-locating cases (n = 11899) were distributed over 2960 unique buildings. Among those co-locating cases, half of them (n = 6466) were co-locating with at most 5 cases in the same building and 1734 cases were found in buildings that were also accommodating over 10 cases (Table 4.2).

Table 4.2 Case distribution in unique building

TB cases were spread in 8355 unique buildings in Hong Kong. One-third of the cases were not co-located with other TB cases in the same building.

Number of cases in a building	Number of building	Percentage of building	Number of TB case	Percentage of TB case
1	5395	64.6	5395	31.2
2-5	2342	28.0	6466	37.4
6-10	490	5.9	3699	21.4
11-20	122	1.5	1562	9.0
over 20	6	0.1	172	1.0
Total	8355	100	17294	100

4.1.2 Aggregated level

TB cases were found in every DCCA. The average number of TB case per DCCA was 43. The highest number of TB cases (n = 103) was found in DCCA *Kwai Tsing Estate* while the lowest number of TB case (n = 6) was found in DCCA *Lamma and Po Toi Islands*. The distribution of TB case by sex and age per DCCA are summarized in Table 4.3 and Table 4.4 respectively. Despite more male TB cases were observed on average, 12.3% (n = 49) of DCCAs were found to have more female cases (Table 4.3). Moreover, the male-to-female ratio varied greatly across 400 DCCAs, with the ratio ranging from 0.1 up to 15. Three percent (n = 12) of DCCAs were found to have male cases 4 times higher than female cases.

Table 4.3 Male-to-female ratio of TB cases by DCCA

In general more male cases were reported in Hong Kong. However, 12% of DCCA were found to have a male-to-female ratio less than 1, meaning that more female cases were notified in those DCCAs.

Male-to-female ratio in one DCCA	Number of DCCA	Percentage of DCCA
1 or below	49	12.3
1.1 – 2	203	50.8
2.1 – 3	110	27.5
3.1 – 4	26	6.5
Above 4	12	3.0
Total	400	100.0

Table 4.4 highlights the percentage of old age TB in each DCCA, which was calculated as the percentage number of case aged 65 or above over total number of TB case in a DCCA. The distribution of old age TB varied over Hong Kong. Forty-seven percent (n = 188) of DCCAs were accommodating more than 40% of cases that belonged to old age TB.

Table 4.4 Distribution of old age TB by DCCA

This table shows the percentage of old age TB over all TB cases in a particular DCCA. Referencing to the global percentage of old age TB in Hong Kong, which is 40%, 47% of DCCA were found to have a higher percentage of old age TB than the global average.

Proportion of old age TB in one DCCA (%)	Number of DCCA	Percentage of DCCA
0 – 10	6	1.5
11 – 20	24	6.0
21 – 30	68	17.0
31 – 40	114	28.5
41 – 50	120	30.0
51 – 60	53	13.3
61 – 70	12	3.0
71 – 80	2	0.5
81 – 90	1	0.3

The variation in TB distribution was likely to be attributed by the underlying population structure, therefore crude notification rate of TB was age-and-sex standardized, giving a standardized notification ratio (SNR). Table 4.5 shows a list of DCCA having a SNR equal or greater than 1.5, meaning that the SNR was 50% higher than the expected. The highest SNR was found in DCCA *Mei Foo North* (SNR = 2.31) in Shum Shui Po District, followed by *Peng Chau and Hei Ling Chau* (SNR = 2.26) in Islands District and *Choi Yuen* (SNR = 1.92) in North District (Table 4.5). Crude notification rate (Figure 4.2) and SNR (Figure 4.3) were also mapped to provide a complete picture. Both maps showed substantial spatial heterogeneity across Hong Kong. A quartile map (Figure 4.2), grouping data at an interval of 25%, reflected the extent of TB notification rate without standardization. Compared to Figure 4.2, the map of SNR (Figure 4.3) exhibited a more specific spatial distribution of high/low TB notification in relation to the expected rate. Respectively 29 and 15 of 400 DCCAs had a SNR 50% higher and 50% lower than expected. DCCAs with higher SNR were mainly located in Kowloon while pockets of high TB were scattered in some new towns in the New Territories and along the coastal areas of Hong Kong Island.

Table 4.5 List of DCCA having TB SNR higher than the expected

District	DCCA	TB SNR	Standardized notification rate	Crude notification rate	Number of TB case
Sham Shui Po	Mei Foo North	2.31	194.43	198.9	91
Islands	Peng Chau & Hei Ling Chau	2.26	189.48	209.7	40
North	Choi Yuen	1.92	161.50	182.5	81
Yau Tsim Mong	Jordan	1.90	159.63	196.3	95
Southern	Stanley & Shek O	1.89	159.10	143.6	87
Kowloon City	Lung Shing	1.81	152.35	178.6	72
Yau Tsim Mong	Mong Kok West	1.78	149.88	175.8	98
Kwun Tong	Kwun Tong Central	1.74	146.07	164.5	70
Sham Shui Po	Nam Cheong Central	1.74	146.01	150.7	82
Wong Tai Sin	Tung Tau	1.74	146.00	200.3	66
Sham Shui Po	Nam Cheong West	1.71	144.09	166.9	63
Kowloon City	Ma Hang Chung	1.71	143.70	141.6	76
Wong Tai Sin	Tsz Wan West	1.69	142.10	149.9	91
Yuen Long	Tsz Yau	1.67	139.99	119.8	81
Kwun Tong	Hing Tin	1.66	139.19	139.6	53
Wan Chai	Hennessy	1.64	137.54	146.3	60
Sham Shui Po	Nam Cheong North	1.63	137.14	148.1	74
Yau Tsim Mong	Sycamore	1.63	136.93	129.3	85
Sham Shui Po	Nam Cheong South	1.62	136.03	152.1	79
Yau Tsim Mong	Mong Kok North	1.61	135.37	146.9	61
Kwai Tsing	Kwai Shing East Estate	1.61	134.86	144.7	85
Wong Tai Sin	Lung Tsui	1.60	134.25	175.6	77
Yuen Long	Tai Kiu	1.58	132.69	121.3	67
Kowloon City	Hung Hom	1.57	131.69	143.9	63
Kwun Tong	Kai Yip	1.55	130.57	189.8	71
Yau Tsim Mong	Yau Ma Tei	1.54	129.38	137.2	68
Tuen Mun	King Hing	1.53	128.73	133.9	68
Kwun Tong	Lok Wah South	1.52	127.88	160.5	58
Sham Shui Po	Nam Cheong East	1.50	126.04	139.0	69
Tuen Mun	Fu Tai	1.50	125.77	98.2	58

Figure 4.2 Map showing TB crude notification rate

Mapping notification rate without adjusting for the age and sex of the underlying population structure allowed a preliminary exploration of TB distribution.

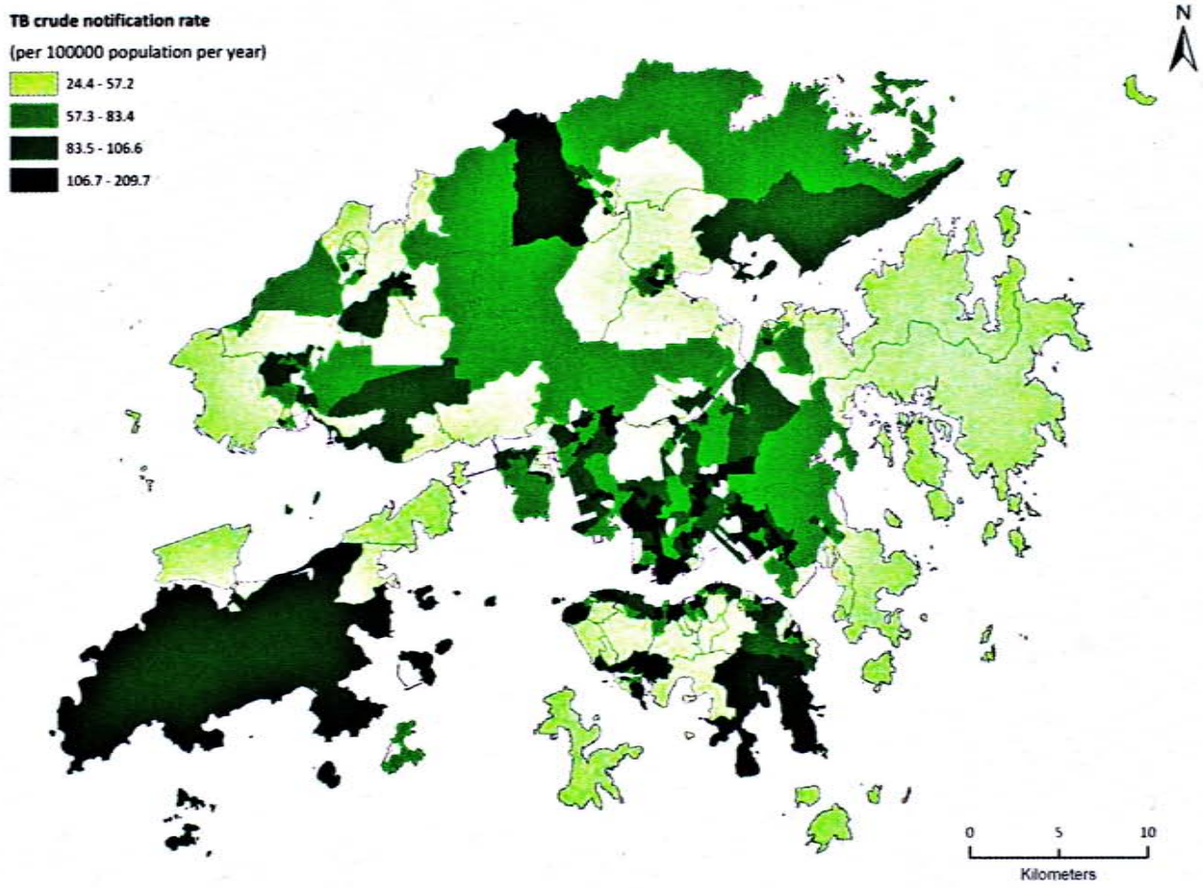
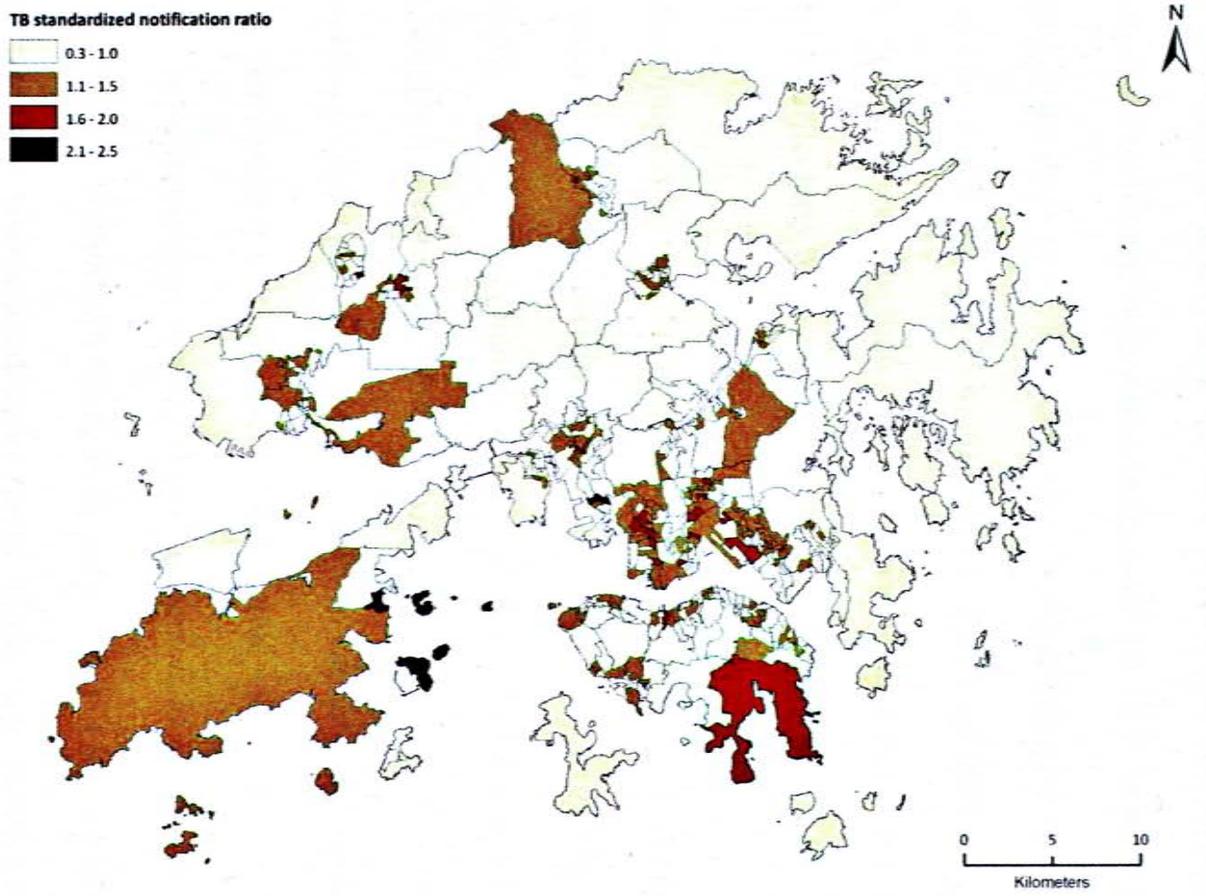


Figure 4.3 Map showing TB standardization notification ratio (SNR)

The SNR map highlighted DCCAs with a particularly higher and lower TB rate than expected, which was more informative than the map of crude notification rate.



To compare the discrepancy with and without standardization, SNR was multiplied by the global notification rate to get a standardized notification rate (Table 4.5). Overall, the discrepancy between standardized notification rate and crude notification rate ranged from -59.25 to 29.69, with an average difference of -2.5 in a DCCA. The negative sign of discrepancy indicated that rate after standardization in general was smaller than the rate before standardization. For DCCAs having SNR greater than 1.5, the discrepancy between standardized notification rate and crude notification rate ranges from -59.25 to 20.16, with an average difference of -14.3 in a DCCA. Both discrepancies indicated that standardization did alleviate the confounding effects particularly for those DCCA with high TB SNR.

A summary of descriptive statistics of neighbourhood determinants is shown in Table 4.6. All variables were expressed in percentage, except *room shared by person* and *population density*. The range of values in a DCCA could be as small as 0.23 to 9.93 for variable s*hort duration of stay* or could be as large as 0 to 100 for variable of p*opulation living in public housing*. For the two variables that could not be expressed in percentage, they were represented in absolute value and interpretation must be with caution. For r*oom shared by person*, as a surrogate for living condition, higher value implied a better living condition. For *population density*, as a surrogate for neighbourhood environment, higher value meant a more crowded living neighbourhood.

Table 4.6 Descriptive statistics of neighbourhood determinants by DCCA

Since most of the neighbourhood determinants were census variables, data values were highly skewed as shown in the table.

Variables	Minimum	Maximum	Mean	Std. Deviation
Non Hong Kong born (%)	21.19	67.04	39.61	8.64
Short duration of residence (%)	0.23	9.93	1.90	1.29
Economically inactive population (%)	34.80	68.33	50.96	5.68
Low income individual (%)	12.09	40.72	22.39	4.94
Population engaging in secondary sector (%)	4.66	27.76	16.60	4.34
Low income household (%)	3.84	54.36	21.33	9.27
Not married (%)	32.64	52.65	41.60	3.04
Small family household (%)	68.43	94.48	83.30	4.31
Large household size (%)	0.96	36.68	13.85	5.45
Room shared by person (number of room per person)	0.48	1.62	1.08	0.23
Crowded quarter (%)	0.94	36.79	14.10	5.34
Population living in public housing (%)	0.00	100.00	47.85	42.52
Population density (number of person per km^2)	152.48	327649.58	73661.20	59418.90
Building coverage (%)	0.32	62.47	19.86	12.50

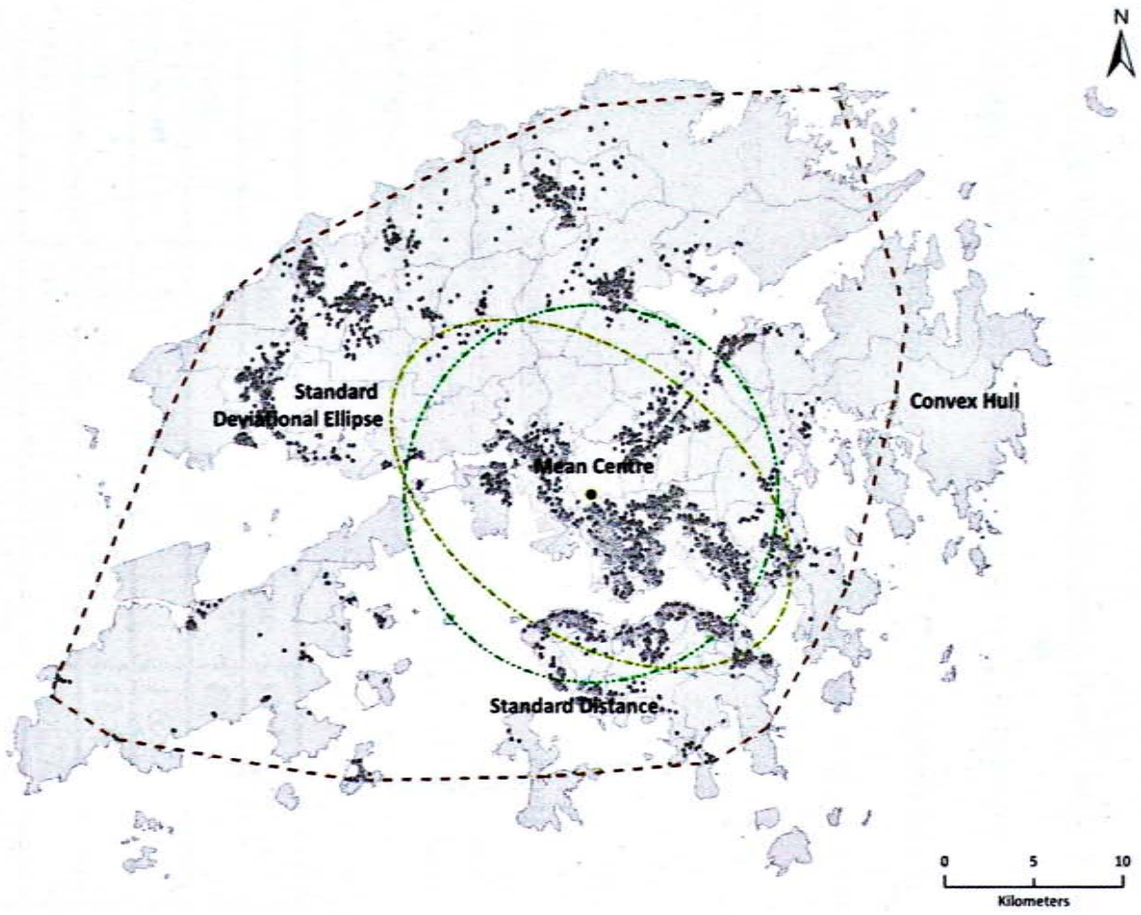
4.2 Results for centrographic analysis

Centrographic analysis was performed in CrimeStat Ver3.2a (Ned Levine, 2009). Point TB cases were used measure the extent, central tendency and dispersion of the spatial distribution. Figure 4.4 shows the centrographic measures of TB cases. In order to protect the spatial confidentiality and prevent possible reverse geocoding, exact location of TB cases were randomly distributed 100 metres away from the exact location in all directions.

In Figure 4.4, the extent of spatial distribution was depicted by the boundary of convex hull as it almost enclosed the whole territory of Hong Kong. The mean centre, as an overall description of central focal point of the data, was located right in the middle of Hong Kong somewhere in the Shum Shui Po District. Both standard distance and standard deviational ellipse were useful tools to illustrate the spatial dispersion. Both of them were used in conjunction with the mean centre to indicate the focal centre of data series and to provide an indication of the distribution of the data around the mean centre. The standard distance and standard deviational ellipse (SDE) was displayed as a circle and an ellipse respectively. Both dispersion measures in Figure 4.4 illustrated spatial dispersion at one standard deviation. Clearly shown in Figure 4.4, the inclination of SDE, comparing to the standard distance, implied that more cases were found in the Northwest and Southeast of Hong Kong. Visual inspection of the data showed that the SDE was skewed in such direction did not only due to the presence of cases in Northwest and Southeast but also the absence of cases in Northeast and Southwest of Hong Kong.

Figure 4.4 Centrographic measures of TB point distribution

Convex hull, mean centre, SDE and standard distance were common indicators of a spatial pattern. As shown in the map, most of the cases were located in the centre of Hong Kong and the distribution slightly followed a Northeast to Southwest direction. (Note: the distribution of TB cases were geomasked to ensure spatial confidentiality)



Using the same dataset, SDE were produced for three different age groups including old age TB (cases aged 65 or above), adult TB (cases aged from 16 to 64) and child TB (cases aged below 16), to show the relative difference in spatial dispersion (Figure 4.5). The SDEs for TB of all age groups followed the general pattern of Hong Kong. The ratio of long to short axis (Table 4.7) was an indicator of the shape of SDE. Larger the ratio meant a more elongated shape and linear distribution of the points. While the ratio gave an indication of distribution circularity, the size of SDE described the compactness of TB concentration. It appeared that the distribution of elderly cases was more clustered than the other age groups, as shown by the smallest size of SDE.

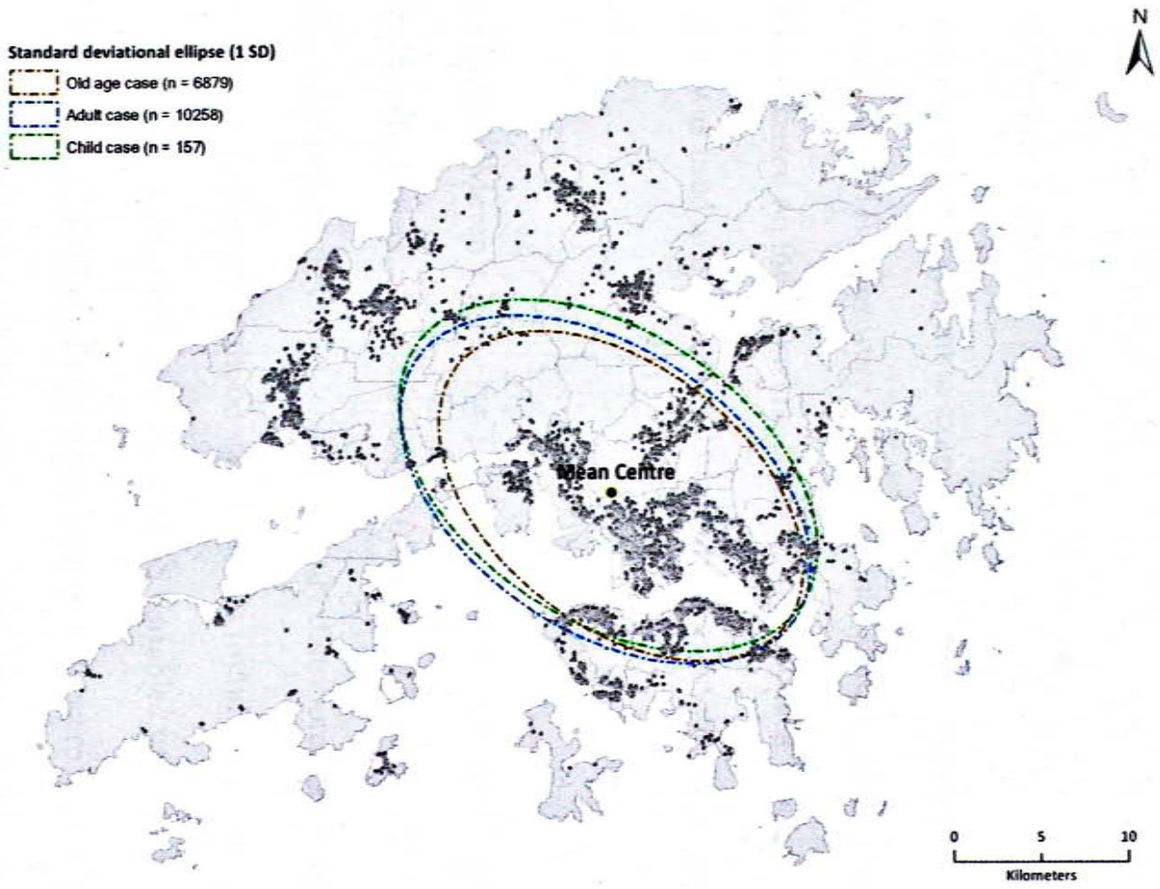
Table 4.7 Table showing the results for SDE (1 Standard Deviation)

Area of the SDE was a good indicator to describe the compactness of spatial distribution. By comparing the size of SDE, the spatial distribution of old age case were more compact than the other age groups

	Child case	Adult case	Old age case
Number of TB case	157	10258	6879
Ratio of long to short axis	1.61	1.63	1.62
Y axis length (m)	16479	16049	14729
X axis length (m)	26506	26186	23877
Area of ellipse (sq m)	343047144	330077157	276208818

Figure 4.5 Map showing the SDE for TB cases in different age groups

The size of SDE for old age TB were smaller than that for other age groups, meaning the spatial distribution of old age TB cases were more compact.



4.3 Results for exploratory spatial data analysis

Two measures of spatial autocorrelation were applied in the ESDA in this study. In this section, results of Moran's Index and Local indicator of spatial association are reported.

4.3.1 Results for Moran's Index

Moran's index is a global measurement to examine the presence of spatial autocorrelation. Individual TB cases were aggregated into DCCA and Moran's Index was calculated based on the ranked neighbourhood variables. Rook contiguity criterion was used to define the relationship of a variable between neighbouring areas.

Table 4.8 shows specific Moran's Index for TB SNR and neighbourhood variables. All Moran Indexes were statistically significant at 0.001 level, meaning the null hypothesis of observed pattern of values being randomly distributed could be rejected. The distribution of neighbourhood variables was found to be spatially clustered at different degree. Moran's Indexes of all 15 variables were positive, indicating that DCCAs with similar values, either high or low, were located near to each other. Larger value of Moran's index indicates a stronger magnitude of the clustering effect. Among all the variables, *population engaging secondary sector* demonstrated the greatest clustering effect (Moran's Index = 0.60). Variables including *building density* (Moran's Index = 0.36), *population non Hong Kong born* (Moran's Index = 0.35), *population living in public housing* (Moran's Index =

0.34) and *small family household* (Moran's Index = 0.31) also demonstrated a relatively high Moran's Index, implying these variables possessed greater spatial clustering effect than other variables.

Table 4.8 Moran's Index of neighbourhood factors and TB SNR

Higher value of Moran's Index implied a stronger clustering effect. Among the neighbourhood variables, population engaging in secondary sector demonstrated the highest clustering effect, which was 3 times greater than that of TB SNR.

Neighbourhood variables	Moran's Index
Non Hong Kong born	0.3482
Short duration of residence	0.2868
Economically inactive population	0.2467
Low income individual	0.098
Population engaging in secondary sector	0.6003
Low income household	0.2496
Not married	0.1448
Small family household	0.3057
Large household size	0.2012
Room shared by person	0.2746
Crowded quarter	0.1814
Population living in public housing	0.3418
Population density	0.2528
Building coverage	0.3607
TB SNR	0.2046

All Moran's index is statistically significant at 0.001 level

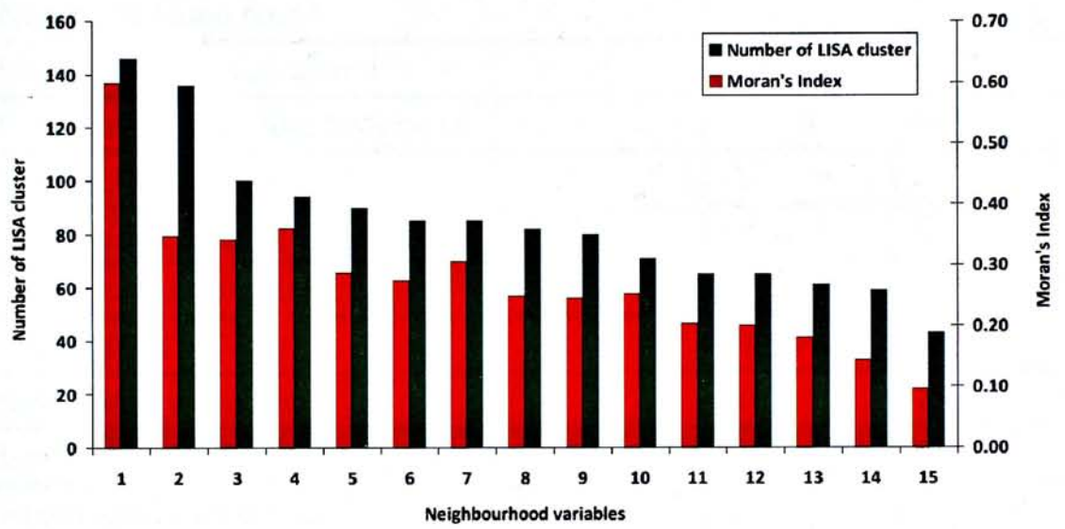
Despite all neighbourhood variables were significantly and positively autocorrelated, a rather weak spatial autocorrelation was observed for the variable of *low income individual* (Moran's Index = 0.098). The low Moran's Index for this variable indicated that there was minimal clustering effect of either high or low value at DCCA level.

4.3.2 Results for Local Indicator of Spatial Association

Since Moran's index is a global measure of spatial autocorrelation, it is not possible to tell if the clustering of high value or low value were attributable to the Moran's index. LISA was subsequently used to localize specific clusters and determine the magnitude of spatial autocorrelation in local level. Four types of local cluster could be indentified by LISA. A High-High cluster means DCCAs of high value is surrounded by DCCAs with high value while a Low-Low cluster means DCCAs of low value is surrounded by DCCAs with low value. A Low-High cluster means DCCAs of low value is surrounded by DCCAs with high value and a High-Low cluster means DCCAs of high value is surrounded by DCCAs with low value. Since Low-high clusters and High-Low clusters are usually treated as spatial outliers, investigation of spatial pattern usually focuses on the High-High cluster and Low-Low cluster which is regarded as hotspot and coldspot respectively. Figure 4.6 depicted the relationship between Moran's Index and number of clusters identified in LISA. Variables having a greater Moran's Index also had a greater number of cluster indentified in LISA.

Figure 4.6 Graph showing the trend of Moran's Index in its relationship with number of LISA clusters identified

Number of local clusters identified by LISA was associated with the value of Moran's Index. *Population engaging in secondary sector* also bore the highest number of local clusters.



To uncover the effect of hotspot / coldspot in contributing to the global clustering effect, the number of local clusters was tabulated for each neighbourhood variables. In Table 4.9, only three variables, including *non Hong Kong born population*, *short duration of residence* and *room shared by person*, have equal or greater number of hotspot than coldspot. Therefore the Moran's Index for these variables was attributed by the presence of hotspots.

Table 4.9 Number of hotspot and coldspot for all variables

LISA is useful to highlight how local clusters in affecting the Moran's Index where the clusters were located. As a result, Moran's Index for three variables were dominantly influenced by the presence of hotspot.

Neighbourhood variables	Hotspot	Coldspot
Non Hong Kong born*	68	68
Short duration of residence*	70	20
Economically inactive population	33	47
Low income individual	15	28
Population engaging in secondary sector	60	86
Low income household	34	48
Not married	19	40
Small family household	40	45
Large household size	26	39
Room shared by person*	62	23
Crowded quarter	22	39
Population living in public housing	26	74
Population density	7	64
Building coverage	35	59
Standardized notification ratio	31	34

To visualize the LISA result, maps of significant cluster from selected variables were generated. All clusters identified were significant at 0.05 level. Hotspots were illustrated in red colour while coldspots were illustrated in blue colour. Figure 4.7 is a cluster map of non Hong Kong born population. Distinctive high concentration was observed in Kowloon area and some parts in Hong Kong Island, while most areas in the New Territories were identified as coldspot. Figure 4.8 shows the cluster of low income household.

Figure 4.7 Map showing significant clusters of non Hong Kong born population

Hotspot was observed in the central part of Hong Kong and coldspots were mostly scattered in the New Territories.

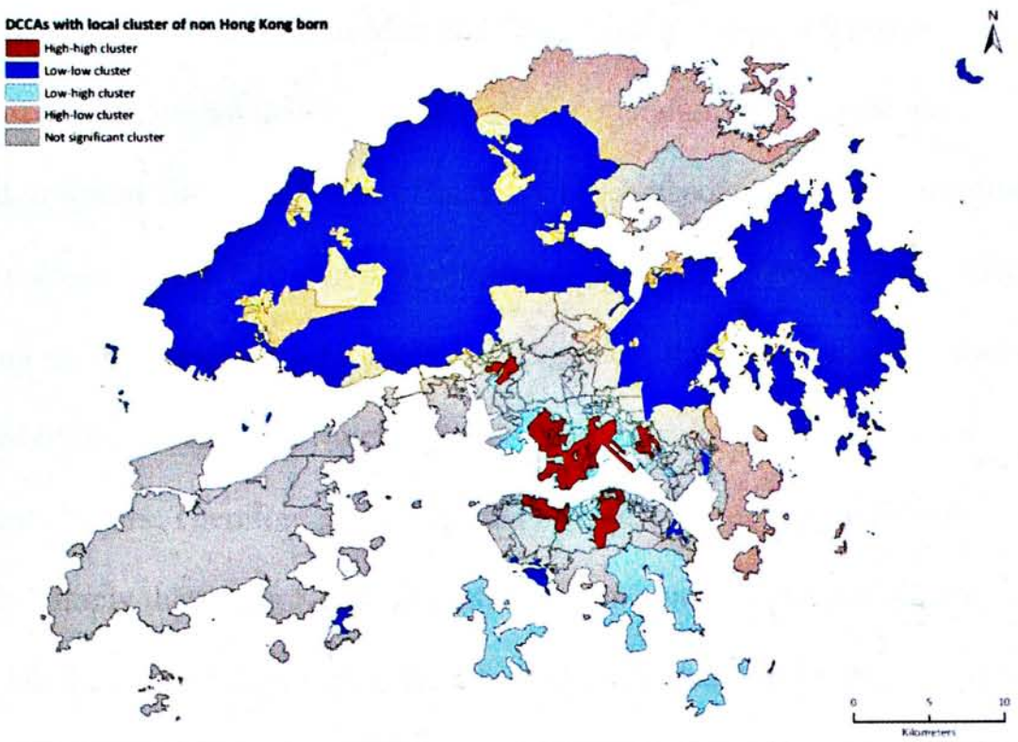
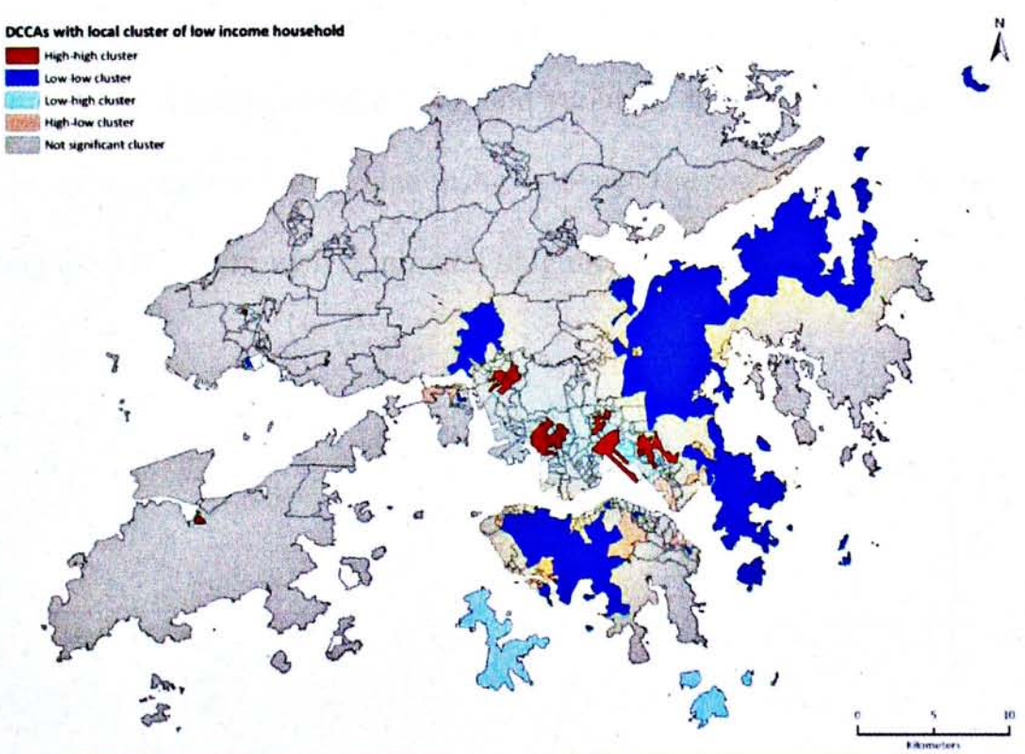


Figure 4.8 Map showing significant clusters of low income household

Most of the significant clusters detected were located in the Eastern and Southern part of Hong Kong. Only few clusters were identified in the Northwestern region.



Most of the significant clusters detected were located in the Eastern and Southern part of Hong Kong. While most hotspots were found in Kowloon area, scattered hotspots were also found in Tuen Mun and Tung Chung. Figure 4.9 shows the clusters of population not married. Coldspots of population not married were observed in Kwun Tong and Eastern District, while hotspots were found in some new town areas. Figure 4.10 shows the cluster of population density. A majority of significant clusters detected were coldspots. However, it should not be overlooked that, most of the coastal area, which are usually densely population, were not significant clusters. Therefore a huge number of coldspots identified did not necessarily imply that Hong Kong was actually dominated by low population density, but the fact was that LISA failed to detect population density as significant clusters. Figure 4.11 shows the distribution of TB SNR cluster. Significant hotspots were found in Kowloon area and coldspot were mostly located in DCCAs with less population.

The spatially varied nature of neighbourhood variables and TB SNR aroused interest in testing if there is any relationship between the spatial variability in explaining the TB pattern with significant neighbourhood variables.

Figure 4.9 Map showing significant clusters of population not married

Coldspot of population not married were found in Kwun Tong and Eastern while hotspots were found in some new town areas.

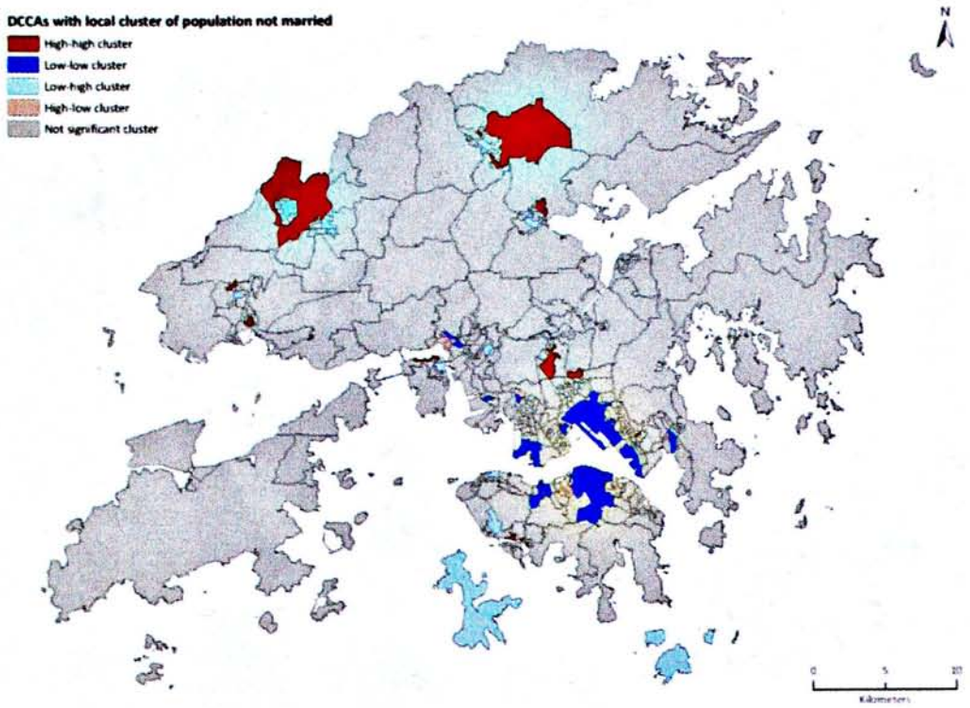


Figure 4.10 Map showing significant clusters of population density

A majority of DCCAs were identified as significant coldspots for population density, meaning that DCCAs with a relatively low population density were located near to each other.

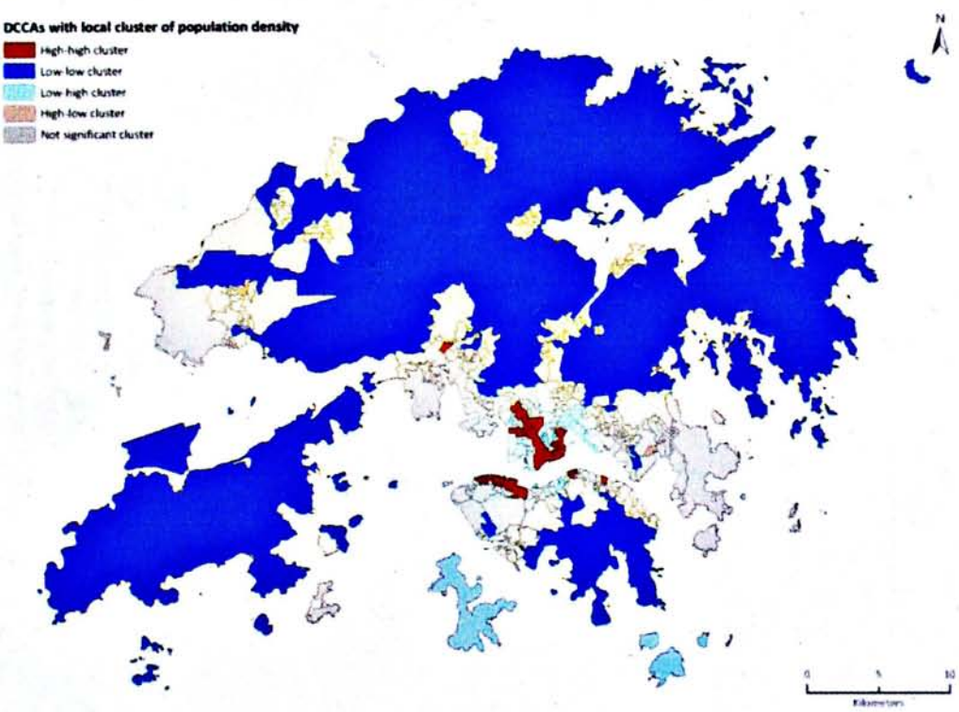
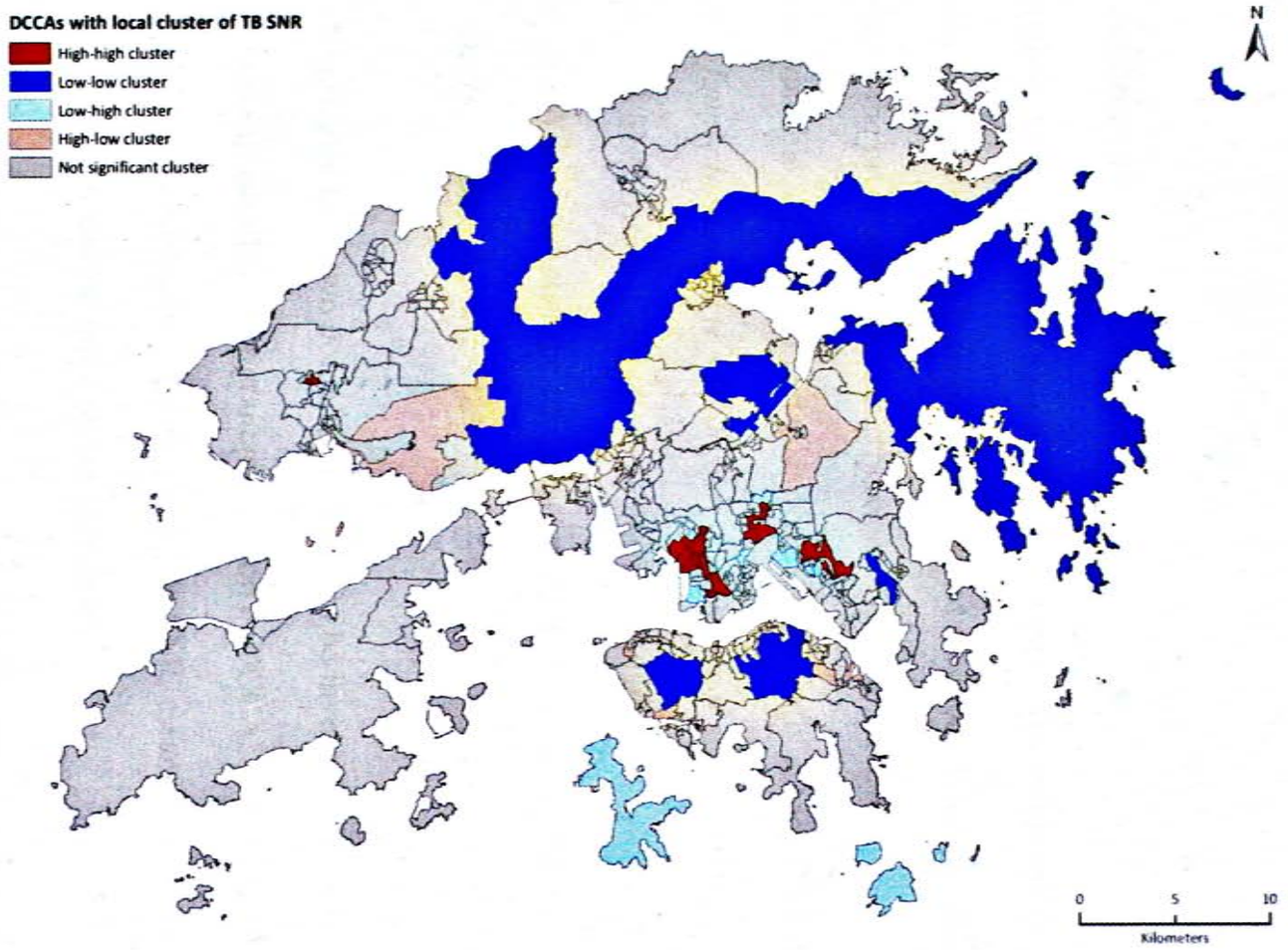


Figure 4.11 Map showing significant clusters of TB SNR

Significant hotspots were found in Kowloon area and most coldspots were located in DCCAs that were sparsely populated.



4.4 Results for explanatory analysis

In this section, results for univariate analysis, multivariate analysis using ordinary linear regression (OLR) model and geographically weighted regression (GWR) model are summarized.

4.4.1 Correlation analysis and variables selection

Univariate analysis using Spearman's correlation coefficient was employed to explore the degree of association between TB SNR and neighbourhood variables. Table 4.10 shows the bivariate relationship between TB SNR and each neighbourhood variable. All but one variables were significantly associated with TB SNR. *Small family household* ($r = -0.06$, $p = 0.24$) was not significantly correlated with TB SNR. Among all positively correlated variables, *low income household* ($r = 0.56$) demonstrated the strongest strength of association to TB SNR, followed by *non Hong Kong born* ($r = 0.39$).

A total of four variables from two domains are found to be inversely correlated to TB SNR. In domain illustrating immigrant population in DCCA, *short duration of residence* ($r = -0.22$) was inversely correlated with TB SNR, meaning that a lower SNR was found in DCCA having more recent immigrants.

Table 4.10 Univariate analysis showing Spearman's R

Spearman's rank correlation were performed to examine the association between TB SNR (the dependent variable) and the neighbourhood variables

Neighbourhood variables	Coefficient
Non Hong Kong born*	0.39
Short duration of residence*	-0.22
Economically inactive population*	0.44
Low income individual *	0.26
Population engaging in secondary sector*	0.21
Low income household*	0.56
Not married*	0.18
Household of small family structure	-0.06
Large household Size*	-0.23
Room shared by person*	-0.50
Crowded quarter*	-0.20
Population living in public housing*	0.35
Population density*	0.29
Building coverage*	0.22

Dependent variable: TB SNR
*Correlation is significant at the 0.01 level (2-tailed)

Three variables in domain illustrating living condition were inversely correlated with TB SNR. These variables were surrogates for crowded living condition which might accelerate a higher TB rate. Among them, *room shared by person* (r = -0.50) demonstrated the strongest strength of relationship with TB SNR, meaning that lower TB SNR was found in DCCA having greater number of room shared per person in one household. However, unusual results were recorded for variable *large household size* (r = -0.23) and *crowded quarter* (r = -0.20). Though it is difficult to

explain why it occured, the ramification of this finding became an important reference for choosing a suitable surrogate marker for crowded living condition in ecological study.

These variables were further analysed in separate linear regression analyses in order to select one variable having the highest partial correlation in each domain (Table 4.11). As a result, *non Hong Kong born population* (partial correlation coefficient = 0.40) in domain of Immigration status, *low income household* (partial correlation coefficient = 0.39) in domain of Economic status, *population not married* (partial correlation coefficient = 0.20) in domain of Family status, *room shared by person* (partial correlation coefficient = -0.26) in domain of Household crowding and *population density* (partial correlation coefficient = 0.20) in domain of Neighbourhood crowding demonstrated the highest partial correlation in each domain . These neighbourhood variables were subsequently chosen as independent variables in further multivariate analysis using OLR.

Table 4.11 Results of linear regression for domain having more than one variable

Variable having the highest partial correlation in each domain were included in the multivariate analysis using OLR

		Unstandardized		Standardized			Correlations		
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
Immigrant status	(Constant)	167.60	13.70		12.23	0.00			
	Non Hong Kong born*	0.39	0.05	0.39	8.67	0.00	0.39	0.40	0.39
	Short duration of residence	-0.23	0.05	-0.23	-5.01	0.00	-0.22	-0.24	-0.23
Economic status	(Constant)	95.11	12.10		7.86	0.00			
	Economically inactive population	-0.06	0.08	-0.06	-0.69	0.49	0.44	-0.04	-0.03
	Low income individual	-0.03	0.05	-0.03	-0.62	0.54	0.26	-0.03	-0.03
	Population engaging in secondary sector	-0.02	0.05	-0.02	-0.39	0.70	0.21	-0.02	-0.02
	Low income household*	0.63	0.08	0.63	8.34	0.00	0.56	0.39	0.35
Family structure	(Constant)	179.91	14.14		12.72	0.00			
	Not married*	0.19	0.05	0.19	3.86	0.00	0.18	0.19	0.19
	Household of small family structure	-0.09	0.05	-0.09	-1.80	0.07	-0.06	-0.09	-0.09
Living condition	(Constant)	311.93	31.21		9.99	0.00			
	Large household Size	-1.57	0.29	-1.57	-5.36	0.00	-0.23	-0.26	-0.22
	Room shared by person*	-0.44	0.08	-0.44	-5.39	0.00	-0.50	-0.26	-0.22
	Crowded quarter	1.38	0.29	1.38	4.80	0.00	-0.20	0.24	0.20
	Population living in public housing	0.07	0.09	0.07	0.80	0.43	0.35	0.04	0.03
Neighbourhood environment	(Constant)	138.20	11.97		11.55	0.00			
	Population Density*	0.26	0.06	0.26	4.07	0.00	0.29	0.20	0.20
	Building coverage	0.06	0.06	0.06	0.89	0.37	0.22	0.05	0.04

Dependent Variable: TB SNR; * Selected variables for the OLR model

4.4.2 Results for ordinary linear regression

Using a stepwise backward conditional approach, an ordinary linear regression was conducted to examine the relationship between TB SNR and selected neighbourhood variables. *Low income household*, *population density*, *population not married* and *non Hong Kong born population* remained independent predictors of the DCCA-based TB SNR (Table 4.12). The surrogate marker for living condition, *room shared by person*, was excluded in the process. All significant variables were positively correlated with SNR, meaning that DCCA with higher TB SNR was found in DCCA with more household with low income, higher population density, more population not married or more population born outside Hong Kong. Among these variables, *low household income* (coefficient = 0.45) was the most significant factor among all variables. The final model explained 35.1% of the total variance in the TB SNR.

Table 4.12 Results of stepwise multiple linear regression model predicting TB SNR

Four out of 5 variables remained significant predictor of TB SNR, with variable *low household income* demonstrating the strongest association.

	Standardized Coefficients	t	Sig.	Correlations		
				Zero-order	Partial	Part
(Constant)	37.4	2.70	0.01			
Low income household	0.45	9.05	0.00	0.56	0.41	0.37
Population density	0.13	3.08	0.00	0.29	0.15	0.12
Not married	0.12	2.85	0.01	0.18	0.14	0.12
Non Hong Kong born	0.11	2.24	0.03	0.39	0.11	0.09
Dependent Variable: TB SNR; Adjusted R Square = 0.35						

4.4.3 Results for geographically weighted regression

Significant predictors identified in the previous multivariate analysis were put into a GWR model. As noted by the substantial degree of spatial autocorrelation of variables, some of the unexplained variance behind the global regression model might be better portrayed by GWR which considers the spatial non-stationarity of association.

Table 4.13a shows the diagnostic information of both models. The AIC value was slightly reduced from 4770.51 in OLR model to 4764.91 for GWR model. The adjusted R-square was raised from 0.35 in OLR model to 0.42 in GWR model. These results suggested that GWR model performed better than the OLR model in terms of explanatory power. The model improvement was further validated by conducting an ANOVA test. The null hypothesis for the ANOVA is that the GWR model does not improve over the OLR model (Table 4.13b). As a result, the F-value (F-value= 2.43) was larger than the critical F-value, therefore the null hypothesis was rejected and GWR was considered to offer a better model performance over OLR at 0.001 level of significance. The improvement also proved that there were significant spatial variations in the relationships of TB SNR and neighbourhood variables.

Table 4.13a Diagnostic information for OLR and GWR

Model performance could be compared by assessing the AIC value and adjusted R-square. Since GWR had a lower AIC value and a higher adjusted R-square, it could be considered as a better model than OLR in this study.

Diagnostic information	OLR	GWR
Residual sum of squares	3434277.86	2810251.22
Sigma	93.24	88.12
Akaike Information Criterion (AIC)	4770.51	4764.91
Coefficient of Determination	0.36	0.47
Adjusted R-square	0.35	0.42

Table 4.13b ANOVA to test model improvement

ANOVA test was performed to test if the reduction in residuals was statistically significant. GWR was proved to be a better model with a significant reduction in model residuals.

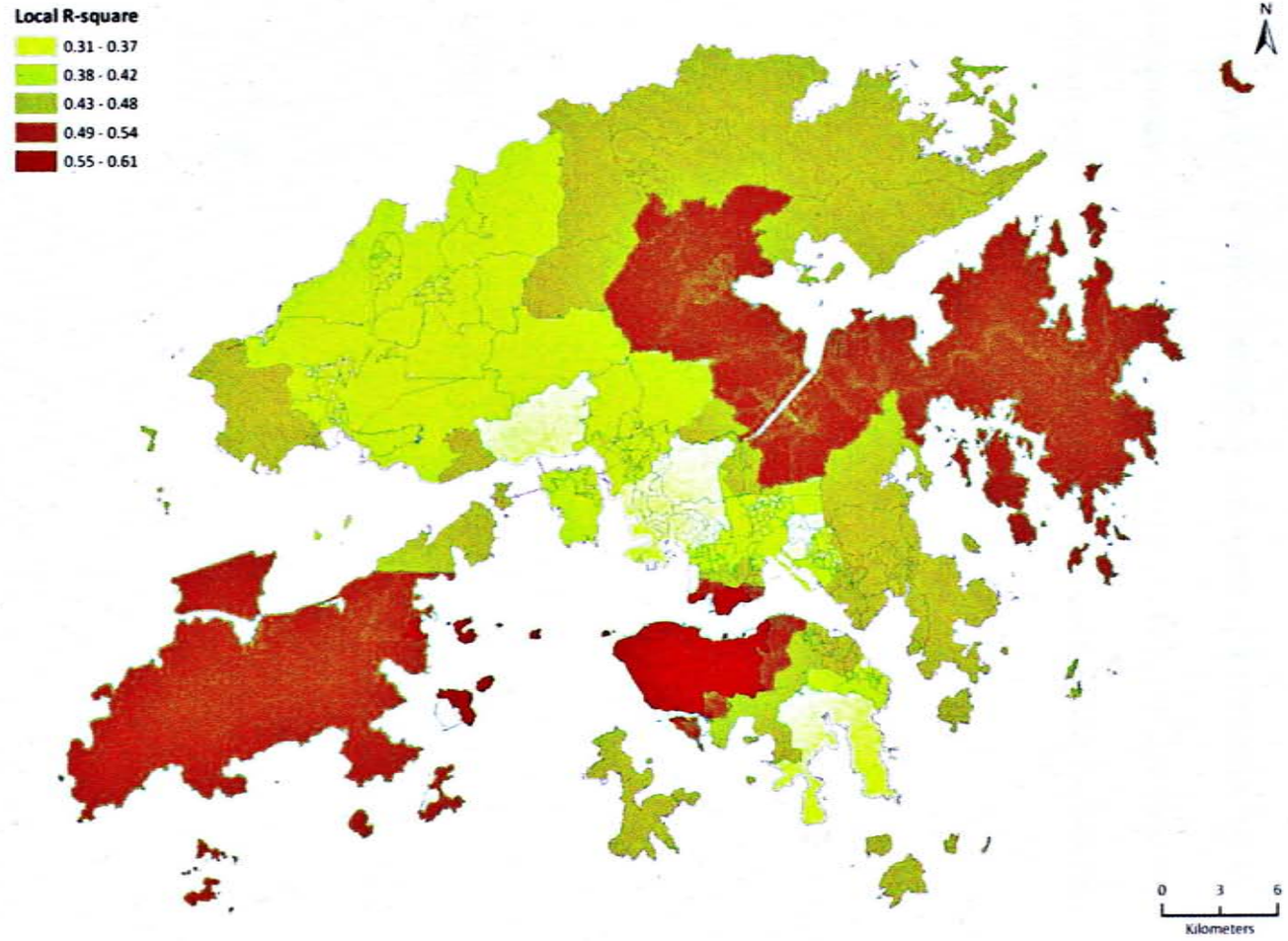
Source	Sum of squares	Degree of freedom	Mean square	F-value	Sig
OLS Residuals	3434277.9	5			
GWR Improvement	624026.6	33.06	18872.73		
GWR Residuals	2810251.2	361.94	7764.52	2.43	0.00

Local R-square for GWR in all DCCA was computed. R-square is particularly informative in understanding the application of the model being calibrated and for exploring the possibility of adding additional explanatory variables to the model. The local R-square of each individual GWR model ranged from 0.31 to 0.61. Using the global R-square ($R^2 = 0.35$) from OLR as threshold, the local R-square for 96% of DCCA (n = 384) in GWR model was greater than that in OLR model. Moreover, 25% of local R-square (n = 100) was greater than 0.5. It thus could be inferred that the relationship between the selected neighbourhood variables and TB SNR was better captured by the GWR model in those regions.

Figure 4.12 shows the distribution of local R-square using a natural break classification method. The local R-square was highest in Wan Chai, Central and Western part of Hong Kong and lowest in the central portion of Kowloon, Southern and most of the Northwestern region of Hong Kong. Despite the generally higher R-square recorded in GWR, it should be also noted that 4% of local R-square was in fact lower than the global r-square predicted in OLR. The spatial variability of local R-square implied the possibilities of location-specific variables in contributing to improvement of the location-specific explanatory power for the model.

Figure 4.12 Map showing the distribution of local R-square from GWR

Natural break data classification was adopted. Illustrated by the map, the explanatory power of GWR was highest in Central and Western District.



In GWR, separate regression equations would be calibrated for each DCCA, therefore resulting in a large volume of output of local parameter estimates and related figures (Appendix 5). Supplemented with the global parameter estimate from OLR, a summary of local parameter estimates of the four variables from GWR is shown in Table 4.14. Since parameter estimates of 4 variables for 400 DCCAs were voluminous, only minimum and maximum value of local parameter estimates was shown as a convenient indication of the extent of the variability.

Table 4.14 Results of local parameter estimates of the variables

Minimum and maximum values of local parameter estimates were compared to the global estimates. As shown in the table, neighbourhood variables in some DCCAs demonstrated an negative association with TB SNR. However, none of the negative estimates were statistically significant.

	Parameter estimates				
Variables	**Local Minimum**	**Global**	**Local Maximum**	**DCCA of different sign as global estimate**	**DCCA of which estimate is significant***
Constant	-13.66	37.4	94.80	2	104
Non Hong Kong born	-0.12	0.11	0.36	76	65
Low income household	0.14	0.45	0.95	0	328
Not married	-0.02	0.12	0.26	7	48
Population density	-0.20	0.13	0.35	109	119

*Parameter estimates were considered significant if the associated t-value is greater than 1.96

Compared to the global OLR estimates, local GWR estimates did reveal spatial variability in the relationship. The wide range of parameter estimates indicated the changing contributions of the variables in the regression equation over Hong Kong (Table 4.14). The change of correlation sign suggested a more complex relationship between TB SNR and neighbourhood variables. In Table 4.14, only variable *low household income* consistently had the same correlation sign as the global parameter estimate. The significance of the parameter estimates is determined by its corresponding local t-values, which is also a product from GWR. In line with results of OLR, DCCA with more low income household was consistently associated with DCCA having higher TB SNR.

The local parameter estimates for other neighbourhood variables vary from negative to positive values. About a quarter of DCCA (n =109) and one-fifth of DCCA (n = 79) were recorded an opposite sign to the global estimation for variable *population density* and *non Hong Kong born population* respectively. However, none of the local estimates having an opposite sign to global estimates for the neighbourhood variables were statistically significant. Among the 4 variables, *low income household* had the greatest number of significant local estimates (n = 328) while *population not married* had the lowest number of significant local estimates (n = 48).

Local parameter estimates were further examined to test for the spatial variability of the relationship between TB SNR and significant variables using Monte Carlo simulation (Table 4.15). As a result, significant spatial variations of the local parameter estimates for two variables, *low income household* and *population density*, were found at 0.05 level of significance, indicating that the varying relationships between the two variables and TB SNR existed at DCCA level. However, the spatial variation for local parameter estimates of the *intercept*, *non Hong Kong born population* and *population not married* were statistically insignificant, meaning that the improvement of GWR model was mainly attributed by the consideration of spatial clustering of *low income household* and *population density*.

Table 4.15 Results of Monte Carlo test for the spatial variability of local parameter estimates

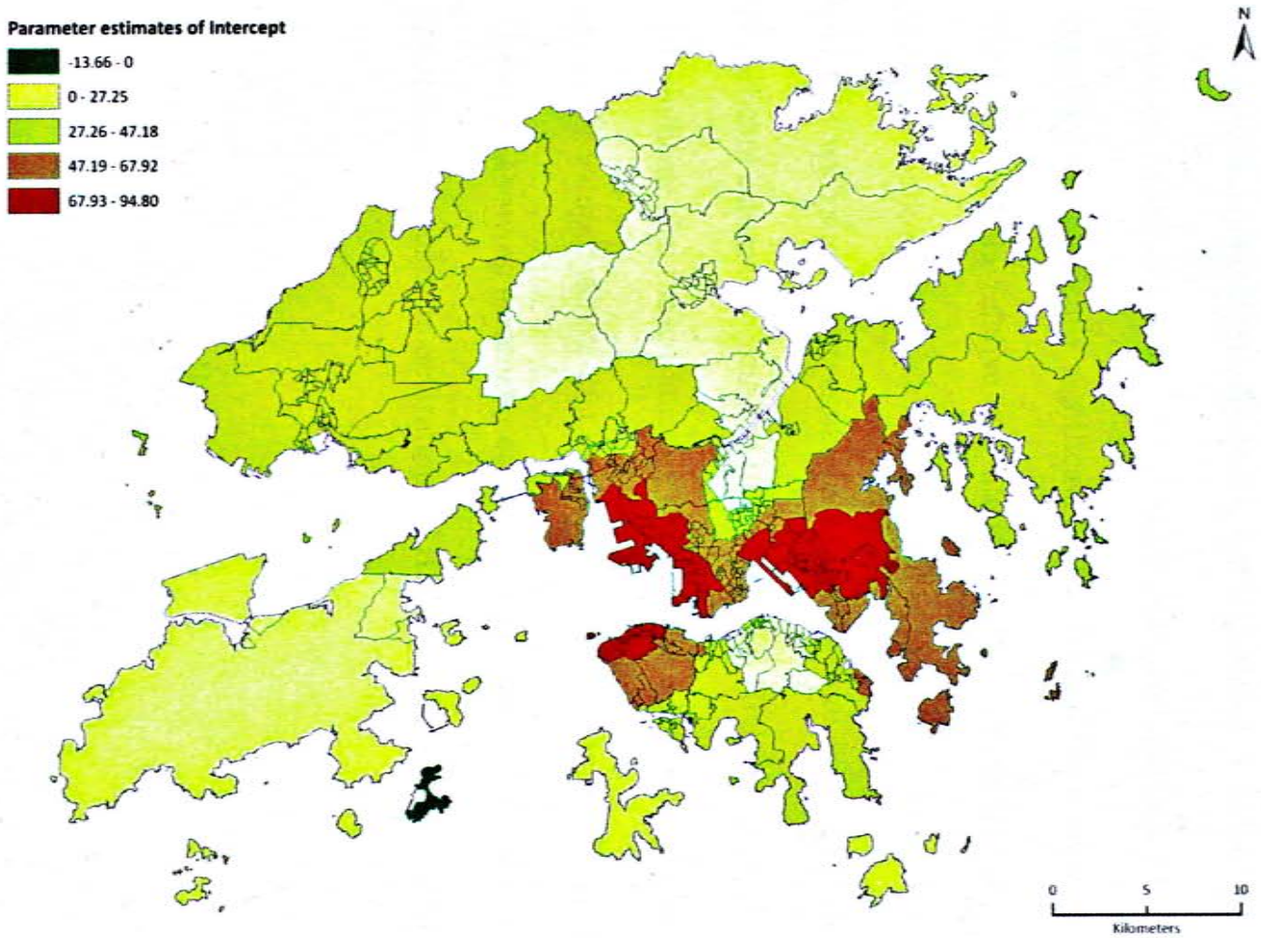
999 Monte Carlo simulations were run to test if the spatial distribution of parameter estimates was significantly varied.

Neighbourhood variables	p-value
Intercept	0.25
Non Hong Kong born	0.15
Low income household	0.02*
Not married	0.94
Population density	0.02*
* Significant at 5% level	

GWR allows mapping of local regression results. The estimates for intercept were depicted in Figure 4.13. Without the influence of other variables, the expected TB SNR was higher in most areas in Kowloon including Yau Tsim Mong, Shum Shui Po, Kwun Tong and Kowloon City, as well as western part of Hong Kong Island. A lower TB SNR was expected in Islands, Wan Chai, Shatin and the northeastern part of Hong Kong. The negative estimates observed in Hei Ling Chau and Ping Chau was however believed to be an outlier.

Figure 4.13 Map showing the parameter estimates of intercept

Without the influence of other variables, the expected TB SNR was higher in most areas in Kowloon as well as western part of Hong Kong Island. The negative estimation in DCCAs *Hei Ling Chau* and *Ping Chau* was believed to be an outlier.



While mapping all the estimate values would provide in general picture of the distribution, mapping only the statistically significant estimates would help highlighting areas that need special attention. A significance threshold (at 95% confidence level) was used to mask out all those areas in which the relationship between the explanatory and dependent variables is not significant. T-values for the 4 statistically significant variables in each DCCA were examined and parameter estimates were classified as statistically significant at 0.05 level if the t-value exceeded 1.96. Maps showing the relative magnitude of significant parameter estimates for each neighbourhood variable are provided in Figures 4.14 to 4.17. For all 4 maps, only DCCA with significant parameter estimates were coloured, while DCCAs with insignificant estimate were filled in grey colour. Higher values of parameter estimates indicate that the explanatory variable has a greater influence in that DCCA where lower values indicate that the predictor variable is less influential in that DCCA. The maps showing parameter estimates highlighted the varying degree of a neighbourhood variable in determining TB SNR in different DCCAs.

Figure 4.14 Map showing the parameter estimates of non Hong Kong born population

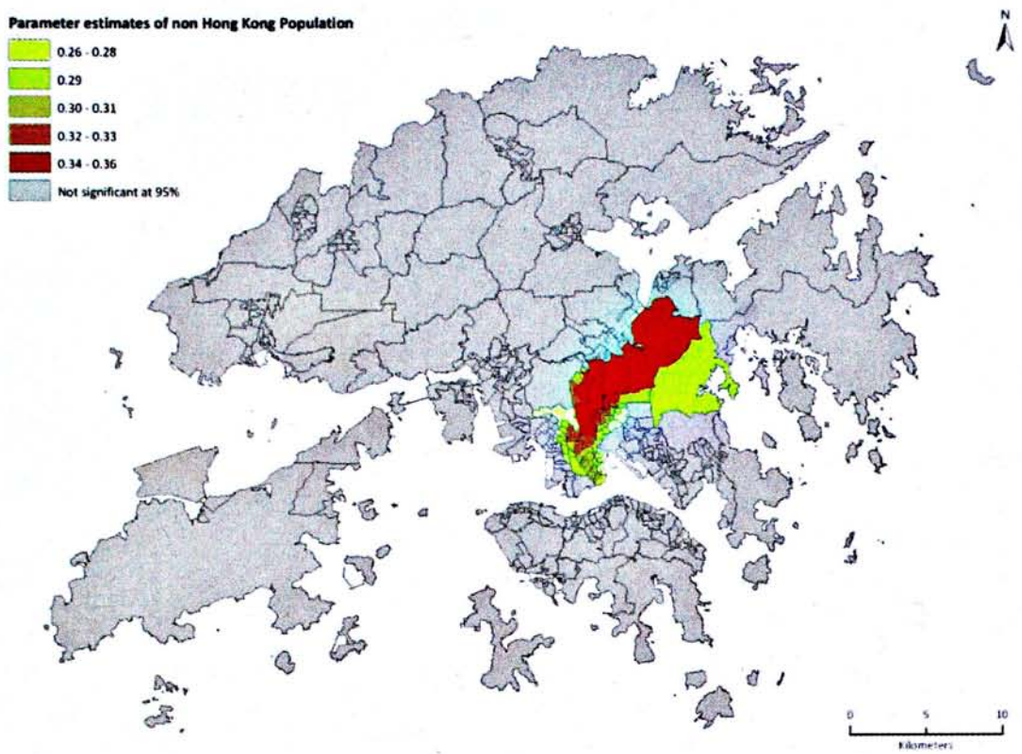


Figure 4.15 Map showing the parameter estimates of low income household

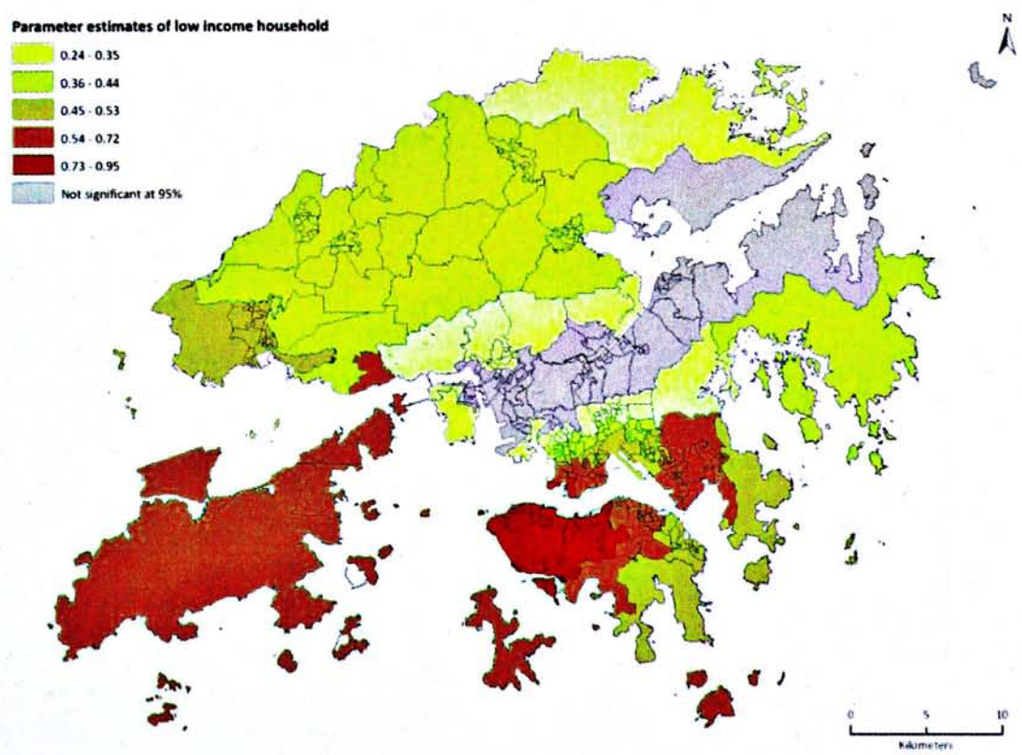


Figure 4.16 Map showing the parameter estimates of population not married

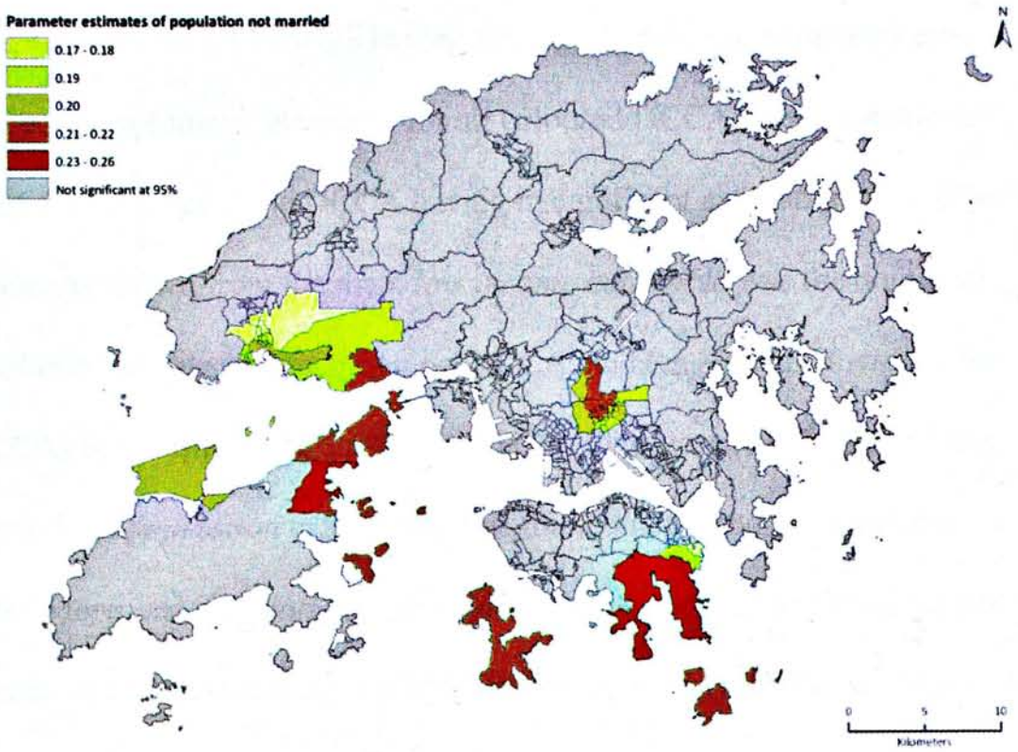


Figure 4.17 Map showing the parameter estimates of population density

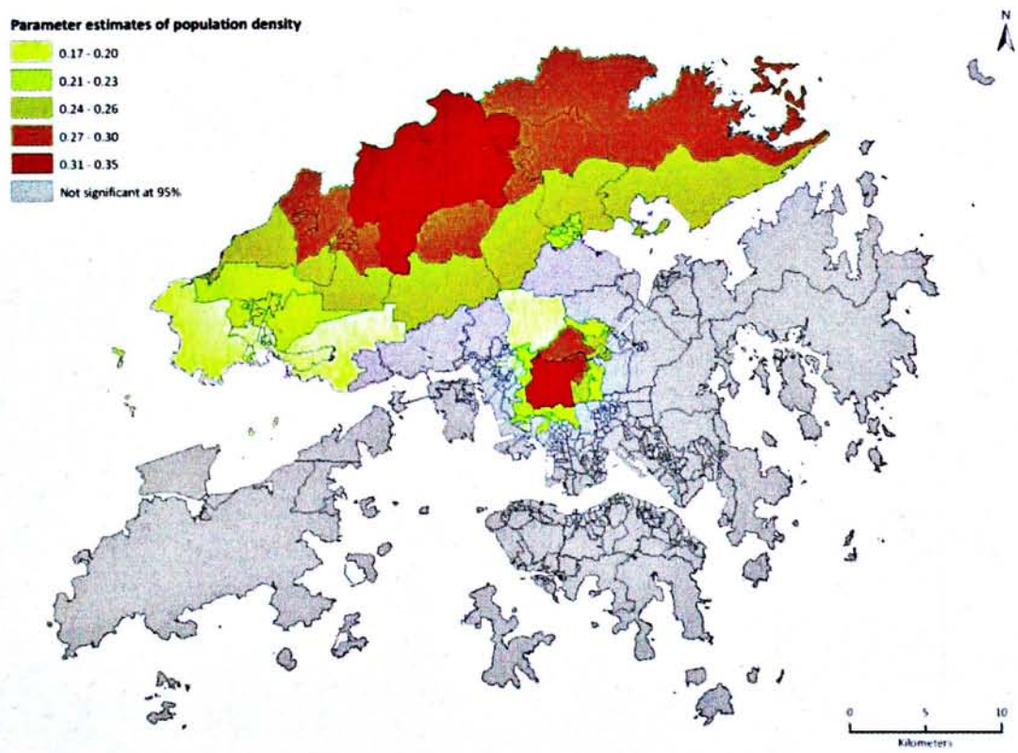
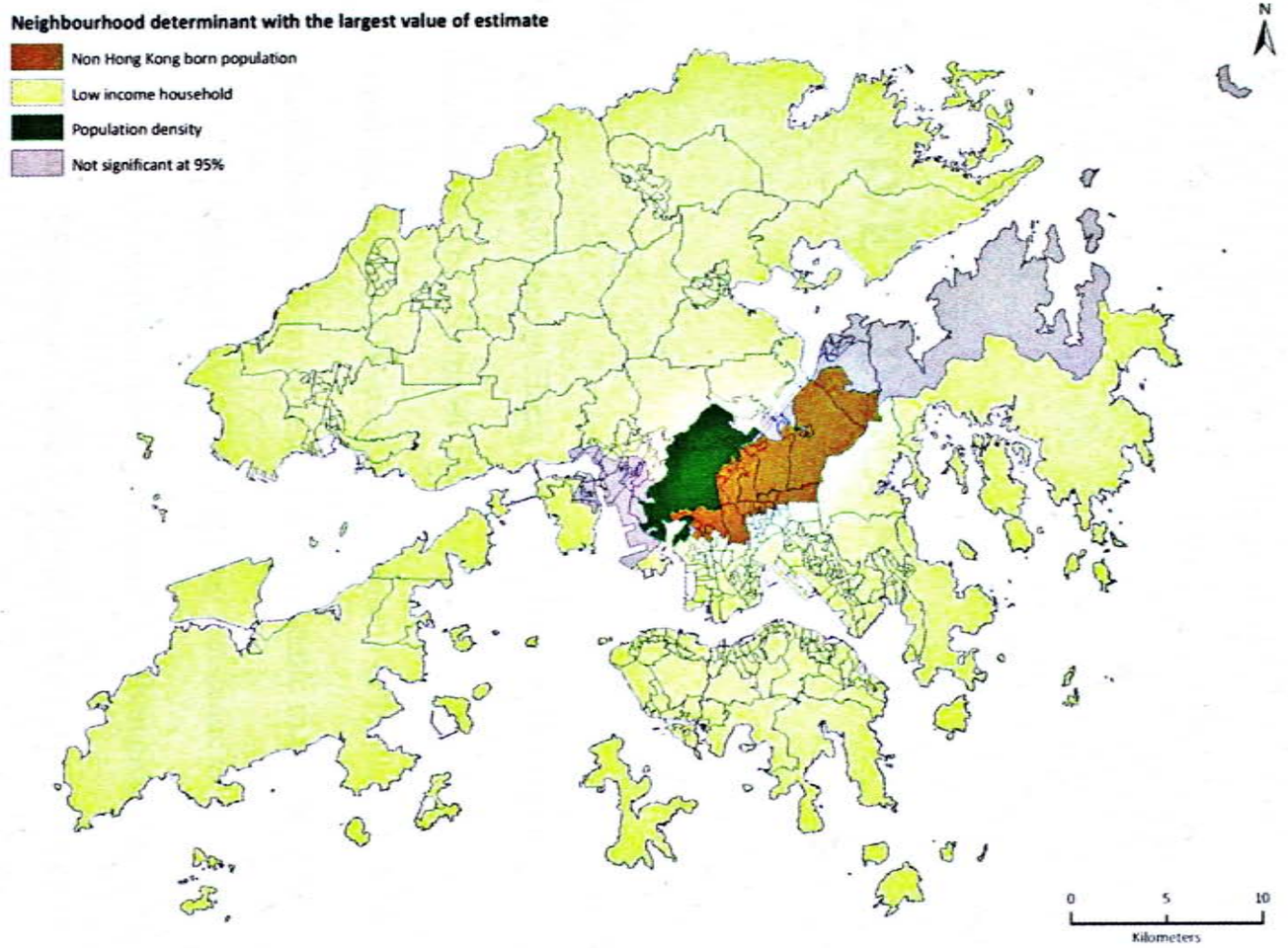


Figure 4.18 was specifically drawn to highlight the dominating variable that posed the greatest influence among variables to TB SNR in a particular DCCA. The map shows the neighbourhood variable having the highest value of parameter estimates in a DCCA. The parameter estimates for all coloured DCCAs were statistically significant at 0.05 level while DCCA having insignificant estimation were filled in grey colour. As shown from the map, *low income household* was the dominant neighbourhood variable in affecting most DCCAs in Hong Kong. However, for some DCCAs in Shatin, Wong Tai Sin and Shum Shui Po, the influence of *non Hong Kong born population* to TB SNR was found greater than *low income household*. Moreover, *population density* also outweighed the magnitude of *low income household* in influencing TB SNR in DCCAs within Shatin and Shum Shui Po District.

Figure 4.18 Map showing the distribution of neighbourhood determinants having the largest parameter estimates

As shown in the map, *low income household* was not always the predominant determinant posing the strongest association to TB SNR.



CHAPTER FIVE DISCUSSION

5.1 Preamble

TB remains a public health problem in Hong Kong, as a region having an intermediate disease burden. The decline of TB in the past 50 years could be attributed to the introduction of anti-TB treatment regimens in 1960s and DOTS in 1970s. However, the decline of TB notification becomes static in recent years. Considering the minimal impact of HIV/AIDS and drug resistance on TB in Hong Kong, non-biomedical factors were examined for possible influence on the local epidemiology. In fact biosocial changes including improvement in socioeconomic status, better housing standard and less malnutrition have collectively led to the decline of TB in industrialized regions (Grange et al, 2009). While a majority of studies examining TB determinants considered the study area as a whole, this study

of spatial epidemiology using an exploratory approach was therefore conducted to bring local variations of TB pattern into focus.

5.1.1 Methods overview

All datasets were georeferenced before performing any analysis. Centrographic measures were used to provide a preliminary exploration of the spatial distribution of TB. ESDA, expanded from the exploratory spatial analysis in conventional statistics, was subsequently used to identify significant global and local clusters of TB and neighbourhood determinants. Possible linkage between cluster of TB and clusters of neighbourhood determinants was assessed by using multivariate analysis. Both OLR and GWR were applied to explore the relationship between neighbourhood variables in different domains and to identify the spatially varying relationship between the significant determinants and TB rate respectively.

5.1.2 Results overview

Short distance between cases and the high number of co-located cases indicated possible clustering among TB cases. Moreover, remarkable heterogeneity in spatial distribution persisted after standardization implied the potentials of neighbourhood determinants that have modulated the TB rates. Moran's Index for all neighbourhood variables and TB SNR were found significantly positive, suggesting the spatial clustering effect of similar values among neighbourhoods. The clustering effect was further validated by LISA, which identified numerous

hotspots, ranging from 7 to 70, in the spatial distribution of TB SNR and neighbourhood determinants.

All neighbourhood determinants but *household of small family structure*, demonstrated a significant univariate association with TB SNR. Among the significant determinants, *low income household* demonstrated the strongest magnitude of association ($r = 0.56$). In order to avoid data multicollinearity, 5 determinants having the highest partial correlation in separate linear regression were selected to a multivariate analysis using OLR. As a result, *low income household*, *population density*, *population not married*, and *population non Hong Kong born* were significant predictors of TB SNR.

Since the OLR model explained only 35.1% of the variance in the TB SNR, other factors influencing TB SNR might have left unaccounted in the model. Based on significant spatial autocorrelation identified through ESDA, it therefore became reasonable to hypothesize that spatial variation in the associations of TB SNR and neighbourhood determinants might constitute the unexplained part in the OLR model.

The increase in R-square, the significant reduction in residual sum of squares and the minimized AIC collectively showed that GWR provided a prediction model with a higher explanatory power than OLR. The significant improvement did not only reflect a better model fitting by GWR in this study but also underscored the

existence and influence of spatial variation on the association between neighbourhood determinants and TB SNR.

5.1.3 Layout of this chapter

In this chapter, the mechanism of significant neighbourhood determinants in influencing the TB pattern and implications of spatial variability are discussed in explaining the spatial distribution of TB. Moreover, two controversial issues relating to study design for spatial epidemiology are addressed. They are the validity of spatial proximity in studying an infectious disease and the adoption of spatial units. Methodological concerns including disease mapping and application of GWR would also be discussed. Probable limitations would be noted before concluding the whole study.

5.2 Neighbourhood determinants in relation to TB

Four neighbourhood determinants but *room shared by person* were identified through stepwise selection and elimination in the OLR. DCCA with higher percentage of low income household, higher population density, more population not married and more non Hong Kong born population were found to have a higher rate of disease.

5.2.1 Crowding and tuberculosis

Crowding has long been identified as a significant determinant of TB. However, the measurement of this factor varied over the world, making its relationship to TB highly contextual (Schmidt, 2008). In this study, two crowding measures were examined. *Population density* was used as a surrogate indicating the effect of neighbourhood crowding to TB rate while *room shared by person* as a surrogate for household crowding.

Room space per person was commonly adopted as a proxy of shared air in household. Crowded household would facilitate household transmission if any household member was infectious. However, the role of household crowding in facilitating transmission has been conflicting. While some studies have uncovered a significant association between household crowding and TB rate (Wanyeki et al, 2006; Souza et al, 2007; Baker et al, 2008), the factor was found insignificantly associated with TB in this study, which is also consistent with a local study (Leung et al, 2004) and a study conducted in New York City (Barr et al, 2001).

Population density is another crowding measure in this study. In order to give a more accurate surrogate, the standard measurement of population density was adjusted. With a mountainous landscape, a large proportion of land is uninhabitable in Hong Kong. Measurement of population density over uninhabitable land area may lead to a misspecification of the magnitude of population density. Uninhabitable areas in each DCCA, which was defined as land over 200 meters in altitude, were therefore excluded using spatial analysis. Population density was equivalent to the population per unit of inhabitable land area. As a result, population density was a significant predictor to TB rate. A positive association between population density and TB rate was also reported in a study conducted in Turkey, highlighting the elevated risk of transmission in crowded areas (Tanrikulu et al, 2008). The result is on the other hand different from a previous local study conducted by Chan-yeung et al (2005), which might be attributed to adjusted measurement of population density in this study.

The discrepancy between two crowding measures in this study might imply the relative importance of household crowding and neighbourhood crowding to TB SNR. The results may suggest that population density would promote transmission and spread outside the immediate household at neighbourhood level while household transmission may be less likely to take place. This explanation is justifiable since Hong Kong is a metropolitan city where population mobility and social mixing within community is high. Carter et al highlighted the unrecognized transmission outside household in large cities and stressed the importance to re-evaluate the complex and dispersed transmission pattern (Carter et al, 2009).

5.2.2 Poverty and tuberculosis

TB is disease of poverty. Poverty conditions could facilitate both disease transmission and disease progression (Barr, 2001; Meyers, 2006; Vendramini, 2006; de Alencar Ximenes, 2009). As noted in Lonnroth's study, people under poverty generally have 1) more frequent contact with infectious agents; 2) a higher likelihood to live in crowded housing with poor ventilation; 3) less food security and nutrients, and 4) lower level of awareness to the disease (Lonnroth et al, 2009).

Due to its multidimensional nature, surrogate for poverty varied greatly in ecological studies, while income level is the most common surrogate markers. In this study, low household income was used as a surrogate for poverty and a strong association was found. The same surrogate markers in two local studies demonstrated conflicting results. While it had a significant association in the Large Street Block Group (LSBG) based study (Chan-yeung, 2004), low household income was not found to be significantly associated with TB rate in the District-based study (Leung, 2004). In this study, a new spatial unit was used, and all surrogate for economic conditions were significantly associated with TB at DCCA-level. Low household income had the strongest influence in predicting TB SNR, both within the economic domain and in the overall model. In spite of having a well-developed social security system with virtual elimination of absolute poverty, together with a good health care infrastructure providing free and highly accessible TB services, the profound effect of the neighbourhood poverty on TB in Hong Kong should not be neglected.

5.2.3 Immigrants and tuberculosis

A person's place of birth reflects the underlying process of migration as well as exposure to TB in the past. Therefore in this study, non Hong Kong born was used as a surrogate to immigrant status. Immigrants were commonly reported as a significant determinant to the TB in designation regions (Vos et al, 2004; Bai et al, 2008; Burzynski & Schluger, 2008) .With up to 30-fold differential TB risks between high- and low-prevalence areas around the world, immigrants contribute a higher proportion of TB cases in low-prevalence areas than that in high-prevalence areas. In Hong Kong, having a moderate TB burden, immigrants may be expected to exert a lesser influence. However, in line with previous analysis (Leung, 2004), TB SNR was found to be positively correlated with the non-Hong Kong-born population in this study. The consistency of the observation suggested that immigrants could nevertheless exert considerable influence on TB epidemiology in Hong Kong. Moreover, the influence of immigrant in contributing to the TB in designation regions could be operated through different mechanisms.

Despite this study was not able to distinguish old immigrants from the recent ones, the majority of recent immigrants in Hong Kong have come from China and they only accounted for a small proportion of total TB cases (Department of Health, 2008). It is then hypothesized that immigrants may contribute to TB in Hong Kong via three possibilities. First, the large reservoir of latent infection among those who immigrated during the war period in the 1940s could have interacted with the aging population to impact TB epidemiology in Hong Kong due to disease reactivation. Second, the family connection in China could have encouraged frequent travel

between China and Hong Kong for immigrants, which implies the exposure to disease and infection may occur in China, where country having a high TB burden. Third, a minority of recent immigrants, including imported workers from other higher incidence areas in Southeast Asia, may also have contributed to this impact.

5.2.4 Marital status and tuberculosis

Being unmarried was used as a surrogate to family support. Relatively few studies reported the association between this determinant and TB. In a previous ecological analysis (Leung et al, 2004) as well as a case-control study (Lienhardt et al, 2005), single marital status was found to be associated with a higher TB rate.

Marital status has been hypothesized to be influential to one's perception of stress and the strength in resisting stress through the presence or absence of family support (Slatcher, 2010). While knowing that depression and stress can exert negative effect on the cell-mediated immune system (Prince et al, 2007), it is therefore reasonable to hypothesize that being single implies an absence of family support, which may in turn increase the vulnerability to disease progression in theory at times of psychosocial stress. However, interpretation of marital status in relation to TB should be careful as it is closely related to other socioeconomic characteristics. Considerable confounding effect among these variables makes it hardly possible to differentiate specific influence of marital status to TB in this ecological analysis. Moreover, the negative and positive effect of being single is not mutually exclusive. To what extend the TB rate being influenced by the absence of family support and the presence of family stress should be further

explored. Examination of the direct effect of family structure at the individual level are required to elucidate the exact mechanism underlying this correlation.

These significant determinants were subsequently further put into a GWR, so that spatial variability of the association and the relative importance of these determinants in the context of different DCCAs could be examined.

5.2.5 Implication of local parameter estimates of association

A series of maps showing the parameter estimates with significant t-values were produced to show how the they varied spatially (Figure 4.14 to Figure 4.17). However, the objective of this study is the identification of significant determinants, therefore assessment of how DCCAs were affected by the varying degree of estimation would not be addressed here. On the other hand, the relative magnitude of the neighbourhood determinants in influencing one DCCA's TB pattern would be discussed. Figure 4.18 was devoted to illustrating the neighbourhood determinants having the strongest association in particular DCCAs. From the results of OLR, low income household was found to be the determinants demonstrating the strongest association. However, from the results of GWR, non Hong Kong born population and population density were found to have a greater association than low income household in some DCCAs located in Shatin, Shum Shui Po and Wong Tai Sin.

This result implied hidden mechanism for association between the determinants and TB SNR at DCCA scale. The difference across space could be intrinsically

caused by some contextual issues, like people's behaviour, that give rise to different responses to the same determinants over space. In this study, most DCCAs in Shatin were mostly influenced by neighbourhood determinants that are different from other DCCAs in Hong Kong. There may be possible linkage between population density, non Hong Kong born population and some specific behavioural pattern in Shatin, such as health seeking behaviour, that lead to more disease transmission and higher TB SNR.

However, it should be noted the relatively low value of R-square, despite the significant improvement from 35% in OLR to 42% in GWR. Due to the limitation of study design, R-square for ecological study is usually small. The lack of individual measures limited the explanatory power of this study as most of the effects of individual exposure left unaccounted for. Therefore an application of multilevel analysis taking into account of neighbourhood and individual effect would enhance the overall model prediction. The inclusion of individual may shed light on both contextual (environmental) and compositional (people) influence on TB epidemiology.

5.3 Study design for spatial epidemiology

Controversy over the spatial dependence and its application in studying spatial epidemiology, and the ethical issues of disease mapping are addressed in this section.

5.3.1 Application of spatial dependence in spatial epidemiology

The concept of spatial dependence, which is computationally expressed as spatial autocorrelation, is a key concept in understanding and analysing spatial pattern. This assumption of spatial dependence is based on the first law of geography invoked by Waldo Tobler in 1970. However, controversy over the validity of this concept in examining spatial pattern has persisted in different disciplines, including sociology, ecology, biology and epidemiology.

Most of the debate over the validity of spatial dependence in studying infectious disease focused on how representative the physical distance in this shrinking world was. However, it should be noted that the core concept of spatial dependence emphasizes on how *near* and *related t*hat features are arranged and it is not necessarily expressed in terms of the actual distance (Miller, 2004). For an airborne infectious disease, face-to-face contact is the most fundamental exposure for a transmission to occur. Moreover, person in close proximity to the infectious agent is more likely to be infected than person far away. The spatial dimension is obvious and the application of the first law of geography in studying infectious disease epidemiology should be valid in general (Krieger, 2003).

Application of the concept should however be cautious. Relatedness does not necessarily mean causality. Instead, relatedness implies the possibility of causality which should be assessed by other evidence. While most of the times spatial autocorrelation is neglected or treated as confounding, it is actually informative as it reveals spatial pattern. (Sui, 2004)

With the developments in GIS, conventional epidemiology methods are expanded in terms of their power and techniques in analysing space-time pattern of disease or adverse health outcomes. Moreover, the improved means of collecting spatial data through technologies such as mobile phones, PDAs and GPS have no doubt contributed to the growth of GIS and public health research.

A fundamental step for all spatial analyses is the collection of spatial data. Discussion has focused on whether residential address is a good indicator of people's exposure. However, in the absence of direct, person-specific environment, residential or working address, according to the context, is considered the best available surrogate for the microenvironment to which population are principally exposed. In this study, residential address was used and geocoded. In the process of geocoding, several problems about in data quality were encountered, which lowered the accuracy of geocoding. The problems are: 1) missing geographical coordinates; 2) location errors associated with geographic coordinates; 3) incorrect or incomplete street addresses; 4) and duplicate records. Despite the technological restriction of geocoding in Hong Kong, manual validation of data largely improved the successful rate of geocoding.

5.3.2 Choosing spatial units

Ecological analysis requires aggregation of data, usually in spatial units. It thus becomes possible for study to associate aggregated disease data with neighbourhood determinants that could not be measured at one single location, for example, population density. As discussed in Chapter Three, choosing an appropriate spatial unit for a study is important as it directly affects the spatial resolution, the statistical stability and the compatibility of study. These dimensions could be approximately examined by the indicators suggested in this study, however, other dimension which could not be easily quantified should also be considered.

Suggested by Boscoe & Pickle (2003), a desirable spatial unit for studying spatial epidemiology may also be considered as appropriate if they possess the following characteristics: *compactness*, *familiarity* and *functional relevance*.

Compactness refers to the shape of spatial units. Ensuring shape compactness would enhance the readability of map as disease pattern over regular areas are usually more recognized than that over irregular areas (Walter, 1993). Although there is no common measurement of the compactness, spatial units having a lower perimeter-to-area ratio are regarded as more compact.

Familiarity refers to how well map reader knows about a certain spatial units. Familiarity of spatial unit was regarded as the "pre-existing knowledge to structures" by MacEachren (2004). According to this criterion, familiarity is not

limited to administrative boundaries. Therefore some studies delineated boundaries that based on residents' perception for conducting spatial analysis. Increase in the familiarity of spatial units could help map reader to link the mapped disease pattern to their pre-existing understanding of the spatial units.

Functional relevance refers to whether the spatial units are conceptually useful in illustrating the theme being mapped. For example, country boundary is not used to illustrate local transmission of disease, while census units are not used to show the global distribution of malaria. Therefore, *functional relevance* is in fact the most important criterion for choosing an analytical unit as it directly relates to the scale of spatial investigation.

5.4 Methodological concern in this study

Methodological concern over disease mapping and parameter setting for GWR is addressed in this section.

5.4.1 Concern over disease mapping

Maps are the most frequently used output of spatial analysis. Disease mapping is a historical means to uncover spatial distribution of disease and to foster hypothesis formulation. With computer development, disease mapping have been undergoing dramatic change in terms of its application and methodology. As a result disease maps are now produced in a much easier way serving growing number of functions. However, increasing channels for convenient disease mapping leads to substantial criticisms over the use of inappropriate cartographic techniques (Cromley & Cromley, 1996; Rytkonen, 2004). Using inappropriate cartographic techniques would lower the communication power of maps and increase the errors of interpretation.

Among all the criticisms, critics over statistical representation of mapped data are mostly found. Disease map of case counts or crude rates seriously limited the explanatory power because of the failure in considering the underlying background population structure which is influential to the disease distribution. Therefore, statistical measures, for instance standardized rate and relative risk, are usually regarded as better mapping measures.

Maps representing point location of patients' address raised increasing awareness of the ethical concern to spatial confidentiality. Studies have demonstrated successful reverse geocoding (Brownstein, 2005; Brownstein & Cassa & Mandl, 2006; Curtis & Mills& Leitner, 2006) and exact address of patients could be identified from published maps. To deal with this issue, geomasking could be applied to tackle the problem of breaching confidentiality. Geomasking is a common way encoding the location of a health record that protects the confidentiality of the person as well as ensures valid geographic analyses of data are possible. The most common method to mask health data is through geographical aggregation. A second technique is random perturbation, which was adopted in this study for the point distribution map of TB. (Armstrong & Rushton & Zimmerman, 1996)

In order to protect spatial confidentiality in this study, each TB case in Figure 4.4 and Figure 4.5 was randomly displaced to a building within 100 meters from its original location in any direction. Moreover, random perturbation performs better if points in DCCAs of low building density are displayed by longer distance than that in high density areas, so that risk of address disclosure in low density areas could be equalized. Despite the effectiveness of geomasking in protecting spatial confidentiality, the application of this method may lead to confusion, if readers are not familiar with map reading. The method may mislead reader in associating TB cases with the new location, in which case unwanted attention is drawn to a location with no TB case.

5.4.2 Application of geographically weighted regression

Unlike conventional regression which produces a single estimation of global relationships among the explanatory and dependent variables, GWR generates a list of statistics for each DCCA. The GWR results in this study for the most part were in line with the results of the OLR model. Table 4.4 provides the mean parameter estimate values for each measure and allows some comparisons between the global and local models. A global linear regression model is not able to accurately characterize the relationship between explanatory and dependent variables if dependent variables exhibit a certain degree of spatial non-stationarity. Therefore, as shown in Chapter Four, local parameters estimates with opposite sign to the global estimations were identified in some DCCAs. Though the parameter estimates with opposite sign was not statistically significant, the possibility of identifying an inverse relationship warrant further exploration.

Mapping local statistics is useful to understand the nature of the model misspecification more clearly. Mapping R-square can be particularly informative to explore the magnitude of GWR in adding additional explanatory power to the OLR. As shown by Figure 4.12, the highest of R-square was found in the western part of Hong Kong Island, where the parameter estimates for low household income were highest. However, only significant parameter estimates should be mapped, of which significance level was indicated by a t-value, in order to yield meaningful interpretation of the results. Some researchers produced map of parameter estimates without highlighting its associated significance level (Holt & Lo, 2008, Edwards et al, 2010). This mapping approach could be very misleading

as it visually draws attention to the areas of extreme parameter estimation, regardless of the significance of the estimate (Jeremy, 2006).

The main difference between GWR and OLR is that the former emphasises on local exceptions and the latter searches for global regularities. There are two major advantages to apply GWR in modelling the spatial epidemiology of TB. First, weighting the influences of neighbouring data points coincides with the need to consider the spatial dimension in modelling an infectious disease. Second, the resulting parameter estimates can be mapped to realize local variations, as well as the relative magnitude of association over the study area.

5.5 Limitation of the study

This study is not without limitations. The 'ecologic fallacy' is always a concern for all ecological studies. Although associations were found between neighbourhood determinants and TB SNR in DCCA, it is not possible to conclude that individuals possessing/exposing to those characteristics are at a higher risk of developing TB. However, MacRae (1994) in his study demonstrated the indirect aggregated effect of socioeconomic deprivation on health status and suggested that the criticism over ecologic fallacy may not be always a valid concern.

Same as ecologic fallacy, vulnerability to MAUP is an inevitable problem of all analyses involving spatial units. The formal definition of MAUP refers to "a problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomenon resulting in the generation of artificial spatial patterns (Openshaw, 1984)." Simply speaking, results of disease rate could vary for different spatial units used. If the identified association of the disease rate and independent variables changes with the selection of different spatial units, the reliability of result is called the MAUP. Despite the lack of a single best solution to tackle this problem, it may in fact not be a problem. Spatial analysis is in nature an iterative process. Direction and strength of association between a dependent variable and independent variables may vary depending on which spatial unit is used. In this case, MAUP is treated as a methodological exploration that the main interest does not focus on testing the significance of association between variables. Instead, examination of the change of association and to what extend of spatial scale leading to the changes draws more interest.

GWR was employed in this study to address the spatial variability. Cautious interpretation of GWR results is required. The kernel distance was defined by the minimization of AIC, which is to produce a best-fit model. As a result, significant spatial clustering of association was uncovered. However, it does not necessarily mean that that neighbourhood determinants showing insignificant spatial variability in this model possess no spatial clustering effect. The spatial clustering effect may exist at different spatial scales and it could be assessed through changing the parameter setting for kernel distance. Therefore, more iterative explorations should be done to investigate possible models in explaining the greatest amount of spatial variability for neighbourhood determinants. Moreover, clustering of TB cases may be influenced by the distribution of urban structure in Hong Kong, it is essential for further study to discuss to what extent the clustering of TB distribution is distinguishable from the distribution of urban settlements.

5.6 Conclusion

Understanding non biomedical determinants of TB has been one of the important dimensions of TB control. On top of identifying neighbourhood determinants of TB, spatial variability of the association between TB and neighbourhood determinants was uncovered in this study. Mapping local statistics highlighted local areas of interest, which allows place-specific control measures could be done. However, the mechanisms underlying the association and observed spatial variability remain largely unexplored. Further studies are needed to determine specific pathways of neighbourhood determinants in leading to high TB rate. Moreover, increasing number of studies demonstrated interest in incorporating spatial analysis with other analytical tools, such ah molecular genotyping and social network analysis in the investigation of TB epidemiology. Spatial analysis is expected to be an important component in studying the epidemiology of TB.

Considerable improvement in information technology has accelerated the development of research opportunities in the field of spatial epidemiology. However, it is, still, a research challenge to have expertise that could handle large volume of high dimensional spatial dataset, perform complex spatial statistics tool and calibrate models which requires careful parameterization.

Reference

Alan M. MacEachren. (2004). *How maps work: Representation, visualization, and design* (1st ed.). New York: The Guilford Press.

Andersen, P., & Doherty, T. M. (2005). The success and failure of BCG - implications for a novel tuberculosis vaccine. *Nature Reviews.Microbiology, 3*(8), 656-662.

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine, 18*(5), 497-525.

Bai, K. J., Chiang, C. Y., Lee, C. N., Chang, J. H., Wu, L. C., & Yu, M. C. (2008). Tuberculosis among foreign-born persons in taiwan, 2002-2005. *Journal of the Formosan Medical Association = Taiwan Yi Zhi, 107*(5), 389-395.

Baker, M., Das, D., Venugopal, K., & Howden-Chapman, P. (2008). Tuberculosis associated with household crowding in a developed country. *Journal of Epidemiology and Community Health, 62*(8), 715-721.

Barr, R. G., Diez-Roux, A. V., Knirsch, C. A., & Pablos-Mendez, A. (2001). Neighborhood poverty and the resurgence of tuberculosis in new york city, 1984-1992. *American Journal of Public Health, 91*(9), 1487-1493.

Beyers, N., Gie, R. P., Zietsman, H. L., Kunneke, M., Hauman, J., Tatley, M., et al. (1996). The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high-incidence community. *South African Medical Journal = Suid-Afrikaanse Tydskrif Vir Geneeskunde, 86*(1), 40-1, 44.

Bishai, W. R., Graham, N. M., Harrington, S., Pope, D. S., Hooper, N., Astemborski, J., et al. (1998). Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy. *JAMA : The Journal of the American Medical Association, 280*(19), 1679-1684.

Brownstein, J. S., Cassa, C. A., & Mandl, K. D. (2006). No place to hide--reverse identification of patients from published maps. *The New England Journal of Medicine, 355*(16), 1741-1742.

Brownstein, J. S., Cassa, C. A., Kohane, I. S., & Mandl, K. D. (2005). Reverse geocoding: Concerns about patient confidentiality in the display of geospatial health data. *AMIA ...Annual Symposium Proceedings / AMIA Symposium.AMIA Symposium,* , 905.

Brownstein, J. S., Cassa, C. A., Kohane, I. S., & Mandl, K. D. (2006). An unsupervised classification method for inferring original case locations from low-resolution disease maps. *International Journal of Health Geographics, 5*, 56.

Brownstein, J. S., Freifeld, C. C., Chan, E. H., Keller, M., Sonricker, A. L., Mekaru, S. R., et al. (2010). Information technology and global surveillance of cases of 2009 H1N1 influenza. *The New England Journal of Medicine, 362*(18), 1731-1735.

Buescher PA. (1997). *Statistical primer: Problems with rates based on small numbers*. North Carolina: Department of Health and Human Services.

Carter, A., Zwerling, A., Olson, S., Tannenbaum, T., & Schwartzman, K. (2009). Tuberculosis and the city. *Health & Place, 15*(3), 807-813.

Chan-yeung, M., Yeh, A. G., Tam, C. M., Kam, K. M., Leung, C. C., Yew, W. W., et al. (2005). Socio-demographic and geographic indicators and distribution of tuberculosis in hong kong: A spatial analysis. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union Against Tuberculosis and Lung Disease, 9*(12), 1320-1326.

Cromley, E. K., & Cromley, R. G. (1996). An analysis of alternative classification schemes for medical atlas mapping. *European Journal of Cancer (Oxford, England : 1990), 32A*(9), 1551-1559.

Curtis, A. J., Mills, J. W., & Leitner, M. (2006). Spatial confidentiality and GIS: Re-engineering mortality locations from published maps about hurricane katrina. *International Journal of Health Geographics, 5*, 44.

Daniel Z. Sui. (2004). Tobler's first law of geography: A big idea for a small world?. *Annals of the Association of American Geographers, 94*(2), 269-277.

Davidow, A. L., Marmor, M., & Alcabes, P. (1997). Geographic diversity in tuberculosis trends and directly observed therapy, new york city, 1991 to 1994. *American Journal of Respiratory and Critical Care Medicine, 156*(5), 1495-1500.

de Alencar Ximenes, R. A., de Fatima Pessoa Militao de Albuquerque,M., Souza, W. V., Montarroyos, U. R., Diniz, G. T., Luna, C. F., et al. (2009). Is it better to be rich in a poor area or poor in a rich area? A multilevel analysis of a case-control study of social determinants of tuberculosis. *International Journal of Epidemiology, 38*(5), 1285-1296.

de Vries, G., Baars, H. W., Sebek, M. M., van Hest, N. A., & Richardus, J. H. (2008). Transmission classification model to determine place and time of infection of tuberculosis cases in an urban area. *Journal of Clinical Microbiology, 46*(12), 3924-3930.

Department of Health. (2008). *Annual report of TB & chest service*. Hong Kong: Department of Health.

Department of Health. (2009). *TB/HIV registry 2009*. Hong Kong: Department of Health.

Donoghue, H. D., Spigelman, M., Greenblatt, C. L., Lev-Maor, G., Bar-Gal, G. K., Matheson, C., et al. (2004). Tuberculosis: From prehistory to robert koch, as revealed by ancient DNA. *The Lancet Infectious Diseases, 4*(9), 584-592.

Dye, C., Lonnroth, K., Jaramillo, E., Williams, B. G., & Raviglione, M. (2009). Trends in tuberculosis incidence and their determinants in 134 countries. *Bulletin of the World Health Organization, 87*(9), 683-691.

Edwards, K. L., Clarke, G. P., Ransley, J. K., & Cade, J. (2010). The neighbourhood matters: Studying exposures relevant to childhood obesity and the policy implications in leeds, UK. *Journal of Epidemiology and Community Health, 64*(3), 194-201.

Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives, 112*(9), 998-1006.

Fotheringham, A. S., Brunsdon C., & Charlton M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships.* England: John Wiley & Sons Ltd.

Francis P. Boscoe, & Linda W. Pickle. (2003). Choosing geographic units for choropleth rate maps, with an emphasis on public health applications. *Cartography and Geographic Information Science, 30*(3), 237-248.

Froggatt, K. (1985). Tuberculosis: Spatial and demographic incidence in bradford, 1980-2. *Journal of Epidemiology and Community Health, 39*(1), 20-26.

Gatrell, A. C., & Bailey, T. C. (1996). Interactive spatial data analysis in medical geography. *Social Science & Medicine (1982), 42*(6), 843-855.

Grange, J. M., Kapata, N., Chanda, D., Mwaba, P., & Zumla, A. (2009). The biosocial dynamics of tuberculosis. *Tropical Medicine & International Health : TM & IH, 14*(2), 124-130.

Haase, I., Olson, S., Behr, M. A., Wanyeki, I., Thibert, L., Scott, A., et al. (2007). Use of geographic and genotyping tools to characterise tuberculosis transmission in montreal. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union Against Tuberculosis and Lung Disease, 11*(6), 632-638.

Harvey J. Miller. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers, 94*(2), 284-289.

Heimer, R., Barbour, R., Shaboltas, A. V., Hoffman, I. F., & Kozlov, A. P. (2008). Spatial distribution of HIV prevalence and incidence among injection drugs users in st petersburg: Implications for HIV transmission. *AIDS (London, England), 22*(1), 123-130.

Higgs, B. W., Mohtashemi, M., Grinsdale, J., & Kawamura, L. M. (2007). Early detection of tuberculosis outbreaks among the san francisco homeless: Trade-offs between spatial resolution and temporal scale. *PloS One, 2*(12), e1284.

Holt, J. B., & Lo, C. P. (2008). The geography of mortality in the atlanta metropolitan area. *Computers, Environment and Urban Systems, 32*(2), 149-164.

Hudelson, P. (1996). Gender differentials in tuberculosis: The role of socio-economic and cultural factors. *Tubercle and Lung Disease : The Official Journal of the International Union Against Tuberculosis and Lung Disease, 77*(5), 391-400.

Jacobson, L. M., de Lourdes Garcia-Garcia, M., Hernandez-Avila, J. E., Cano-Arellano, B., Small, P. M., Sifuentes-Osornio, J., et al. (2005). Changes in the geographical distribution of tuberculosis patients in veracruz, mexico, after reinforcement of a tuberculosis control programme. *Tropical Medicine & International Health : TM & IH, 10*(4), 305-311.

Jansa, J. M., Serrano, J., Cayla, J. A., Vidal, R., Ocana, I., & Espanol, T. (1998). Influence of the human immunodeficiency virus in the incidence of tuberculosis in a cohort of intravenous drug users: Effectiveness of anti-tuberculosis chemoprophylaxis. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union Against Tuberculosis and Lung Disease, 2*(2), 140-146.

Jia, Z. W., Jia, X. W., Liu, Y. X., Dye, C., Chen, F., Chen, C. S., et al. (2008). Spatial analysis of tuberculosis cases in migrants and permanent residents, beijing, 2000-2006. *Emerging Infectious Diseases, 14*(9), 1413-1419.

Kistemann, T., Munzinger, A., & Dangendorf, F. (2002). Spatial patterns of tuberculosis incidence in cologne (germany). *Social Science & Medicine (1982), 55*(1), 7-19.

Krieger, N. (2003). Place, space, and health: GIS and epidemiology. *Epidemiology (Cambridge, Mass.), 14*(4), 384-385.

Lai, P. C., So, F. M. & Chan, K. W. (2009). S*patial Epidemiological Approaches in Disease Mapping and Analysis*. New York: CRC Press.

Lai, P. C., Wong, C. M., Hedley, A. J., Lo, S. V., Leung, P. Y., Kong, J., et al. (2004). Understanding the spatial clustering of severe acute respiratory syndrome (SARS) in hong kong. *Environmental Health Perspectives, 112*(15), 1550-1556.

Leung, C. C., Yew, W. W., Chan, T. Y., Tam, C. M., Chan, C. Y., Chan, C. K., et al. (2005). Seasonal pattern of tuberculosis in hong kong. *International Journal of Epidemiology, 34*(4), 924-930.

Leung, C. C., Yew, W. W., Tam, C. M., Chan, C. K., Chang, K. C., Law, W. S., et al. (2004). Socio-economic factors and tuberculosis: A district-based ecological analysis in hong kong. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union Against Tuberculosis and Lung Disease, 8*(8), 958-964.

Lienhardt, C. (2001). From exposure to disease: The role of environmental factors in susceptibility to and development of tuberculosis. *Epidemiologic Reviews, 23*(2), 288-301.

Lofy, K. H., McElroy, P. D., Lake, L., Cowan, L. S., Diem, L. A., Goldberg, S. V., et al. (2006). Outbreak of tuberculosis in a homeless population involving multiple sites of transmission. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union Against Tuberculosis and Lung Disease, 10*(6), 683-689.

Lonnroth, K., Jaramillo, E., Williams, B. G., Dye, C., & Raviglione, M. (2009). Drivers of tuberculosis epidemics: The role of risk factors and social determinants. *Social Science & Medicine (1982), 68*(12), 2240-2246.

Luquero, F. J., Sanchez-Padilla, E., Simon-Soria, F., Eiros, J. M., & Golub, J. E. (2008). Trend and seasonality of tuberculosis in spain, 1996-2004. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union Against Tuberculosis and Lung Disease, 12*(2), 221-224.

MacRae, K. (1994). Socioeconomic deprivation and health and the ecological fallacy. *BMJ (Clinical Research Ed.), 309*(6967), 1478-1479.

Mennis, J. (2006). Mapping the results of geographically weighted regression. *The Cartographic Journal, 43*(2), 171-179.

Mitchison, D. A. (2005). The diagnosis and therapy of tuberculosis during the past 100 years. *American Journal of Respiratory and Critical Care Medicine, 171*(7), 699-706.

Moreno, S., Baraia-Etxaburu, J., Bouza, E., Parras, F., Perez-Tascon, M., Miralles, P., et al. (1993). Risk for developing tuberculosis among anergic patients infected with HIV. *Annals of Internal Medicine, 119*(3), 194-198.

Munch, Z., Van Lill, S. W., Booysen, C. N., Zietsman, H. L., Enarson, D. A., & Beyers, N. (2003). Tuberculosis transmission patterns in a high-incidence area: A spatial analysis. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union Against Tuberculosis and Lung Disease, 7*(3), 271-277.

Murray, J. F. (2004). A century of tuberculosis. *American Journal of Respiratory and Critical Care Medicine, 169*(11), 1181-1186.

Myers, W. P., Westenhouse, J. L., Flood, J., & Riley, L. W. (2006). An ecological study of tuberculosis transmission in california. *American Journal of Public Health, 96*(4), 685-690.

Nerlich, A. G., Haas, C. J., Zink, A., Szeimies, U., & Hagedorn, H. G. (1997). Molecular evidence for tuberculosis in an ancient egyptian mummy. *Lancet, 350*(9088), 1404.

Nunes, C. (2007). Tuberculosis incidence in portugal: Spatiotemporal clustering. *International Journal of Health Geographics, 6*, 30.

Onozuka, D., & Hagihara, A. (2007). Geographic prediction of tuberculosis clusters in fukuoka, japan, using the space-time scan statistic. *BMC Infectious Diseases, 7*, 26.

Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M. R., et al. (2007). No health without mental health. *The Lancet, 370*(9590), 859-877.

Randremanana, R. V., Sabatier, P., Rakotomanana, F., Randriamanantena, A., & Richard, V. (2009). Spatial clustering of pulmonary tuberculosis and impact of the care factors in antananarivo city. *Tropical Medicine & International Health : TM & IH, 14*(4), 429-437.

Rytkonen, M. J. (2004). Not all maps are equal: GIS and spatial analysis in epidemiology. *International Journal of Circumpolar Health, 63*(1), 9-24.

Sakula, A. (1982). Robert koch: Centenary of the discovery of the tubercle bacillus, 1882. *Thorax, 37*(4), 246-251.

Schmidt, C. W. (2008). Linking TB and the environment: An overlooked mitigation strategy. *Environ Health Perspect, 116*(11)

Selwyn, P. A., Sckell, B. M., Alcabes, P., Friedland, G. H., Klein, R. S., & Schoenbaum, E. E. (1992). High risk of active tuberculosis in HIV-infected drug users with cutaneous anergy. *JAMA : The Journal of the American Medical Association, 268*(4), 504-509.

Serpa, J. A., Teeter, L. D., Musser, J. M., & Graviss, E. A. (2009). Tuberculosis disparity between US-born blacks and whites, houston, texas, USA. *Emerging Infectious Diseases, 15*(6), 899-904.

Slatcher R. B. (2010) Marital functioning and physical health: Implications for social and personality psychology. *Social and Personality Psychology Compass 3*, 1-15.

Souza, W. V., Carvalho, M. S., Albuquerque Mde, F., Barcellos, C. C., & Ximenes, R. A. (2007). Tuberculosis in intra-urban settings: A bayesian approach. *Tropical Medicine & International Health : TM & IH, 12*(3), 323-330.

Stead, W. W. (1997). The origin and erratic global spread of tuberculosis. how the past explains the present and is the key to the future. *Clinics in Chest Medicine, 18*(1), 65-77.

Tam, C. M., Leung, C. C., Noertjojo, K., Chan, S. L., & Chan-Yeung, M. (2003). Tuberculosis in hong kong-patient characteristics and treatment outcome. *Hong Kong Medical Journal = Xianggang Yi Xue Za Zhi / Hong Kong Academy of Medicine, 9*(2), 83-90.

Tanrikulu, A. C., Acemoglu, H., Palanci, Y., & Dagli, C. E. (2008). Tuberculosis in turkey: High altitude and other socio-economic risk factors. *Public Health, 122*(6), 613-619.

Tiwari, N., Adhikari, C. M., Tewari, A., & Kandpal, V. (2006). Investigation of geo-spatial hotspots for the occurrence of tuberculosis in almora district, india, using GIS and spatial scan statistic. *International Journal of Health Geographics, 5*, 33.

Uthman, O. A. (2008). Spatial and temporal variations in incidence of tuberculosis in africa, 1991 to 2005. *World Health & Population, 10*(2), 5-15.

Vendramini, S. H., Santos, M. L., Gazetta, C. E., Chiaravalloti-Neto, F., Ruffino-Netto, A., & Villa, T. C. (2006). Tuberculosis risks and socio-economic level: A case study of a city in the brazilian south-east, 1998-2004. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union Against Tuberculosis and Lung Disease, 10*(11), 1231-1235.

Vos, A. M., Meima, A., Verver, S., Looman, C. W., Bos, V., Borgdorff, M. W., et al. (2004). High incidence of pulmonary tuberculosis persists a decade after immigration, the netherlands. *Emerging Infectious Diseases, 10*(4), 736-739.

Vynnycky, E., & Fine, P. E. (1997). The natural history of tuberculosis: The implications of age-dependent risks of disease and the role of reinfection. *Epidemiology and Infection, 119*(2), 183-201.

W.R Tobler. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography, 46*, 234.

Walter SD. (2001).
Disease mapping: A historical perspective. In P. Elliott, Wakefield J., Best N. & Briggs D. (Eds.), *Spatial epidemiology: Methods and applications* (). USA: Oxford University Press.

Walter, S. D. (1993). Visual and statistical assessment of spatial clustering in mapped data. *Statistics in Medicine, 12*(14), 1275-1291.

Wanyeki, I., Olson, S., Brassard, P., Menzies, D., Ross, N., Behr, M., et al. (2006). Dwellings, crowding, and tuberculosis in montreal. *Social Science & Medicine (1982), 63*(2), 501-511.

Weber, D. J., Rutala, W. A., Samsa, G. P., Sarubbi, F. A.,Jr, & King, L. C. (1989). Epidemiology of tuberculosis in north carolina, 1966 to 1986: Analysis of demographic features, geographic variation, AIDS, migrant workers, and site of infection. *Southern Medical Journal, 82*(10), 1204-1214.

Wilson, L. G. (1990). The historical decline of tuberculosis in europe and america: Its causes and significance. *Journal of the History of Medicine and Allied Sciences, 45*(3), 366-396.

World Health Organization. (2005). *Tuberculosis control in south-east asia and western pacific regions 2005 : A biregional report.* Geneva: World Health Organization.

World Health Organization. (2009). *Global tuberculosis control : Epidemiology, strategy, financing : WHO report 2009*. Geneva: World Health Organization,.

World Health Organization. (2010). *Multidrug and extensively drug-resistant TB (M/XDR-TB): 2010 global report on surveillance and response*. Geneva: World Health Organization.

Wu, P., Cowling, B. J., Schooling, C. M., Wong, I. O., Johnston, J. M., Leung, C. C., et al. (2008). Age-period-cohort analysis of tuberculosis notifications in hong kong from 1961 to 2005. *Thorax, 63*(4), 312-316.

Yanagawa, H., Hara, N., Hashimoto, T., Yokoyama, H., & Tachibana, K. (1981). Geographical pattern of tuberculosis and related factors in japan. *Social Science & Medicine.Medical Geography, 15D*(1), 141-148.

Yeh, Y. P., Chang, H. J., Yang, J., Chang, S. H., Suo, J., & Chen, T. H. (2005). Incidence of tuberculosis in mountain areas and surrounding townships: Dose-response relationship by geographic analysis. *Annals of Epidemiology, 15*(7), 526-532.

Appendix

Appendix 1 How to calculate TB SNR?

Based on the distribution of population by age-and-sex in each DCCA and the age-and-sex specific rates of TB in the Hong Kong population, expected number of TB cases is estimated and then compared to the observed number of TB cases in each DCCA. The ratio of observed and expected is called TB standardized notification ratio (TB SNR). DCCA having TB SNR greater than 1 suggests that TB rate is higher than expected, even after adjusting the population structure, while DCCA having a TB SNR lower than 1 means that the TB rate is lower than the expected. In order to be comparable to crude notification rate, a TB Standardized Notification *Rate*, though not commonly used, is also computed by multiply the standardized

notification ratio in each DCCA with the crude notification rate in the Hong Kong population.

Appendix 2 How GWR works?

GWR is a recent refinement of the ordinary linear regression model to analyse spatially varying relationships among variables. Three models including Gaussian, Logistic and Poisson are available in GWR for different data types. The primary assumption of GWR is the spatial autocorrelation among data. Instead of calibrating a single regression equation, GWR incorporates each observation into a separate regression model using a series of distance-related weights. The distance-related weight, usually termed a kernel, defines how the weight of an observation changes when distance varies. Depending on the study design, fixed or adaptive shape of kernel could be used and the bandwidths of kernel may be manually chosen or optimized using an algorithm. With the kernel, parameter estimates at any DCCA (regression point) are no longer depending on its own value only but also on the values of neighbouring DCCAs.

Appendix 3 What is AIC?

Akaike's Information Criterion (AIC) is a measure of the goodness of fit of an estimated statistical model developed by Hirotsugu Akaike. The AIC is an operational way of trading off the complexity of an estimated model against how well the model fits the data. Increasing the number of predictors generally improves the goodness of fit, regardless of the number of actual (true) predictors in the data generating process. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages overfitting. The preferred model is the one with the lowest AIC value. The AIC methodology attempts to find the model that best explains the data with a minimum of predictors. It is a relative measure, so you can't compare the AIC from one study to another, independent study.

Appendix 4 How Monte Carlo test works?

The test is an iterative process that involves randomly rearranging the data to different locations, recalculating parameter estimates and variances, and comparing those variances with the original model's variance (where the data were in the correct location). The result is a *p* value for each parameter that indicates whether significant spatial variation in the relationship between that parameter and the violence measure exists.

Appendix 5 List of GWR output

GWR will be calibrated in every data point. For each data point, a specific set of figures will be generated. Based on these figures local variations of the relationship between dependent variable and independent variable could be mapped. Local statistics include:

1. The local coefficient for variables
2. T values associated with the parameter estimates.
3. The observed value
4. The GWR predicted value
5. The local R-square statistic